Journal of Information Hiding and Multimedia Signal Processing ©2024 ISSN 2073-4212
Ubiquitous International Volume **15**, Number **1**, March **2024**

Enhancing Image Retrieval Efficiency through Text Feedback to Improve Search Performance

Phuc Nguyen[1,2]

[1]Faculty of Information Systems, University of Economics and Law, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
phucnq@uel.edu.vn

Received December 22, 2023; revised February 22, 2024; accepted February 26, 2024

Abstract. *In the realm of e-commerce, the increasing prevalence of diverse customer needs has led to a growing trend in the research and development of alternative search engine approaches. To address the limitations of existing search models, there is a rising interest in more versatile search methods. In recent years, various approaches have emerged that integrate both images and text to cater to the demand for flexible searches. The aim of this model is to generate a product image that not only mirrors the input image but also adjusts its details based on provided text. This involves using a reference image and text feedback as input to achieve the desired outcome.*

Combining low-to-high detail descriptions to best satisfy user requirements is a highly successful method of application in this retrieval model. We experiment with various techniques based on the image retrieval with text feedback model, compare their efficacy on the Shoes dataset, and offer suggestions for improvement. In comparison with the current approach, we primarily concentrated on and obtained results on image feature extraction, text feature extraction, and the optimizer algorithm with +2.98%, +1.14%, and +3.52% Recall@50 accuracy. Along with accuracy, we also evaluate resource loss and model training time when making recommendations for model optimization.

Keywords: Image retrieval, Search performance, Text feedback, Multi-grained uncertainty regularizations.

1. **Introduction.** In contemporary times, the shopping experience has transcended the traditional in-store product selection, owing to the significant expansion of the digital industry. Globally, individuals have become accustomed to engaging in e-commerce. Through e-commerce platforms, businesses of all sizes can establish online stores, providing users with the convenience of browsing, selecting, and comparing a diverse array of products.

The search industry has grown significantly over the past several decades thanks to developments in personalization, natural language processing, and multimedia results. Beginning with the search engine in the 1990s [14], improved techniques gradually emerged alongside newly proposed, more sophisticated search models like image search, voice search, etc.

As human demand increases, the old-fashioned search models no longer satisfy customers when they go to e-commerce websites. Take a simple example: there was a woman named Sarah who loved shopping for clothes online. She was an avid user of e-commerce sites and would spend hours browsing through different products, trying to find the perfect outfit for any occasion. Initially, Sarah relied on text search queries to find

the products she was looking for. She would enter specific keywords such as "black dress" or "leather boots" in the search bar and would get a list of results based on these queries. However, Sarah found that text search queries had several limitations. Sometimes she was unable to find the exact product she was looking for, as the search results were not always accurate or complete. One day, Sarah discovered image search, where she could search for products by uploading an image of the product or a similar item. This allowed her to find more accurate and complete results, as the search algorithm would use the visual features of the product to match it with similar items. Sarah was delighted with this feature, as it helped her find products that were difficult to describe in words.

However, Sarah still encountered some issues with image search. Sometimes the search algorithm would not be able to accurately match the product with similar items, resulting in irrelevant or incomplete results. Additionally, Sarah found that the search algorithm was not always able to capture her preferences accurately, as it lacked context and personalization.

This is where image retrieval with text feedback came in. Sarah discovered that she could provide feedback on the images retrieved by the system, allowing the platform to learn and improve its results over time. She could highlight specific features of the product she was looking for and provide additional context to help the system understand her preferences better. With image retrieval with text feedback, Sarah found that she could find products that more accurately matched her preferences. She could provide feedback

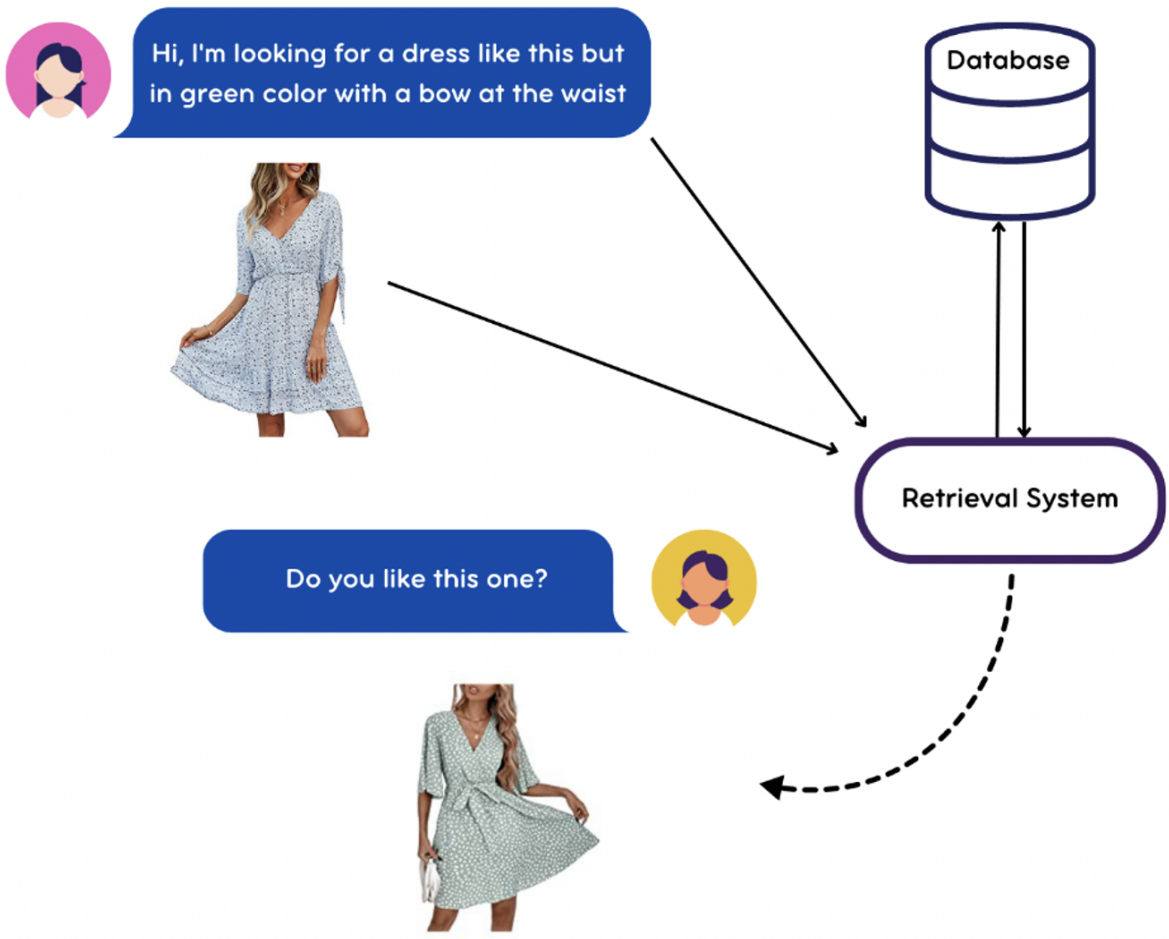

FIGURE 1. Potential application scenario of the image retrieval with text feedback system.

on the images retrieved by the system, resulting in a more personalized and effective search experience.

By providing feedback on the images retrieved by the system, users can help the platform learn and improve its results. By this idea, the image retrieval with text feedback system is illustrated in Fig. 1.

The introduction of image retrieval with text feedback has significantly improved the user experience by providing a more intuitive and accurate way of searching for products, leading to increased customer satisfaction and potentially increased sales for e-commerce platforms. Although this new approach is still not widely recognized, research on it is growing. It has the potential to develop into an extremely powerful search tool for all users, not just those who use e-commerce sites.

2. **Related works.** Deep learning (DL) has made significant strides in numerous fields recently, particularly in computer vision, where machine intelligence has surpassed that of humans. The deep architecture uses a non-linear transformation to integrate low-level features into abstract high-level features, giving it the ability to learn the semantic representation from images.

As a result of the above, it is possible to increase the performance of image retrieval by utilizing the findings of DL on computer vision (CV). Over the past decade, numerous image search models have been developed and refined to improve the accuracy and relevance of search results. Some of the most notable models can be listed here are Convolutional Neural Networks (CNNs) [12] - deep learning models that use convolutional layers to extract hierarchical features from images, or Deep metric learning [7] - a type of machine learning technique that learns a distance metric between images based on their visual similarity.

When the limitations of image search lead to the creation of image retrieval with text feedback model, the use of separate text and image encoders to extract linguistic and visual elements of the query is a common practice in many image search systems. This is because text and image data have inherently different characteristics that require different techniques for feature extraction. Text data can be processed using natural language processing (NLP) techniques to extract semantic and syntactic features, while image data can be processed using computer vision techniques to extract visual features.

Using existing approaches for text and image encoding can be beneficial as it saves the time and resources required to develop custom models. Popular text encoders include BERT [5], GloVe [13], while popular image encoders include VGG, ResNet, and Inception. However, it is important to select the appropriate encoder models for a given task, as different encoders may perform better or worse depending on the specific requirements of the application. Additionally, combining the text and image encodings in a meaningful way is also important for improving the performance of the overall system.

The linguistic and visual elements of the query are typically extracted independently by employing the text encoder and image encoder using the existing approaches [2]. To match the visual features of the target image, these two types of features are combined as the final query embeddings. Depending on whether a pre-trained model is used, there are typically two families of works on image retrieval with text feedback. Recently, there is a composed image and text query model called ComposeAE that was proposed in 2021 to learn the composed representation of image and text query [1]. Using a novel strategy, Muhammad et al. contend that the source image and the target image are located in the same complex space. They are rotations of one another, and query text features are used to convey the degree of rotation. They changed their training aim to include a rotating symmetry loss in response to their innovative articulation of the issue. Their

tests on three datasets reveal that ComposeAE consistently performs better on this task than the SOTA approach [15]. To assure fair comparison and pinpoint its shortcomings, they improve the SOTA approach [16].

However, the composeAE method is tested on a data set called MITstate [8], this dataset has a relatively small size, which may not accurately reflect the complexity and diversity of real-world image retrieval scenarios. This can limit the ability of models trained on these datasets to generalize to new and unseen data. Additionally, the MITstate datasets are limited in terms of the number and diversity of visual and textual features, which can impact the effectiveness of models that rely on these features. Specifically, this data set only gives simple descriptive texts consisting of an adjective and a noun, not properly representing the natural language of the users in reality. It primarily focused on object recognition and classification tasks, and may not be well-suited for more complex image retrieval tasks that require semantic understanding and context.

Focusing more on text issues, Visiolinguistic Attention Learning (VAL) [3], by using an attention mechanism to focus on the most important visual and linguistic features, is able to generate highly informative and accurate representations that can handle a wide range of text feedbacks, including attribute-like and natural language descriptions.In comprehensive tests, VAL has consistently outperformed competing methods across all datasets, demonstrating its superior ability to capture and process visiolinguistic information. This makes it a valuable tool for a wide range of applications, from image retrieval and recommendation systems to natural language processing and computer vision.

An important factor in image retrieval with textual information is how to composite image and text features. Content-Style Modulation (CoSMo) proposes a novel image-based compositor that addresses the challenge of compositing image and text features in image retrieval with textual information [10]. The CosMo model employs two separate modulators to composite image and text features, namely the content modulator and the style modulator. The content modulator first conducts local modifications on the visual feature map using the extracted visual feature from the reference image in accordance with the text features. This approach allows the content modulator to selectively modify certain regions of the visual feature map while preserving the overall structure of the image. The style modulator, on the other hand, learns to generate a style embedding vector from a given text input. This vector is then used to adjust the style of the visual feature map, which is obtained from a pre-trained convolutional neural network (CNN). The combination of the content and style modulators allows CosMo to generate high-quality images that are both content-rich and stylistically diverse. CosMo has shown promising results in image retrieval with textual information, outperforming existing methods in terms of accuracy and efficiency.

The next step is provided by CLVCNet [6], or Cross-Layer Vision and Language Composition Network, which is a novel approach that combines image and text features to improve image retrieval. The model utilizes two fine-grained compositors: a local-wise image-based compositor and a global-wise text-based compositor. These compositors work together, learning from each other while ensuring consistency in their predictions. This mutual learning helps to improve the composition of image and text features both locally and globally, but it still has limitations [17]. One possible limitation is that the mutual learning approach used in the model can be computationally expensive and time-consuming, especially when dealing with large datasets. Another potential disadvantage is that the model heavily relies on pre-trained models, which may not always generalize well to other datasets or domains. Additionally, the model may not be effective in scenarios where the input data contains significant noise or outliers, as the mutual learning approach may amplify errors and inaccuracies in the data.

Based on the CoSMo study [10], Yiyang Chen et al. have studied and proposed a method close to reality called Multi-grained Uncertainty Regularization [4]. This method aims to improve the accuracy of image retrieval by incorporating both coarse-grained and fine-grained matching techniques. By providing the user with several options based on their initial input and allowing them to refine their requirements through feedback, the system can better understand the user's needs and provide more accurate results. The use of uncertainty learning is also interesting as it can help the system avoid missing potential candidates that may match the user's wishes. Uncertainty learning is a technique that allows the model to learn from its mistakes and update its predictions based on the level of confidence it has in its own predictions. Overall, this approach seems like a promising way to improve the accuracy of image retrieval systems and provide users with more relevant results. However, it would be important to evaluate its performance on a large dataset and compare it to other state-of-the-art methods to fully understand its effectiveness.

3. **Experiment process.**

FIGURE 2. Shoes dataset's architecture with 4 main categories with 27 subcategories.

3.1. **Data overview.** Many fashion datasets are currently utilized to train image retrieval models with text. For instance, we have large and well-known datasets like 1) Fashion200k, which is large-scale (more than 200k images) and diverse collection of fashion images it contains. The images of Fashion200k dataset cover a wide range of clothing and the textual information focuses on attribute-like descriptions, and 2) FashionIQ, another large-scale dataset (more than 77k images), which the impressive feature off this dataset is the questions are designed to be complex, requiring the model to have a deep understanding of both the image and the associated textual information. The questions cover a wide range of topics such as color, texture, pattern, and style, as well as more subjective aspects such as outfit coordination and occasion-appropriateness. A dataset called Shoes, which crawls 10,751 pairs of shoe photos from like.com and contains relative expressions that explain fine-grained visual contrasts in images, has also been used in recent studies.

In this work, we will test the model on Shoes dataset because of those reasons such as 1) it contains rich and detailed textual descriptions that can be used to guide the user's search. The additional attributes such as brand and color can also be used to narrow down the search space and improve the accuracy of the retrieval. 2) It satisfies the condition for text feedback that the text must be natural and realistic. This allows for a more natural human-computer connection by allowing users to explicitly convey in normal language the most salient conceptual differences between the preferred search item and the previously retrieved content. 3) The size of this dataset is consistent with the scope of the study since we only have a limited time to do this research. This dataset contains triplets with more opulent properties and fewer label mistakes for dia-logue-based interactive retrieval. For training, we use 10,000 samples, and for evaluating, 4,658 samples. There are 27 subcategories in the Shoes dataset's image data, with bags, earrings, ties, and women's shoes serving as the major categories. The Shoes dataset's architecture is shown in Fig. 2.

The Shoes dataset includes text files (caption) for training and test sets and visual data. The paths for the reference image, the target image, and the caption that catenates between the reference image and the target image will be gathered in these text files (in one line) (see Fig. 3).

data/shoes/attributedata/womens_athletic_shoes/1/img_womens_athletic_shoes_1488.jpg;data/shoes/attributedata/womens_athletic_shoes/0/img_womens_athletic_shoes_ 987.jpg;are gray with green accents

data/shoes/attributedata/womens_stiletto/1/img_womens_stiletto_1596.jpg;data/shoes/attributedata/womens_stiletto/0/img_womens_stiletto_116.jpg;have a gray leopard pattern

data/shoes/attributedata/womens_stiletto/0/img_womens_stiletto_116.jpg;data/shoes/attributedata/womens_stiletto/0/img_womens_stiletto_906.jpg;are matte black with no print

data/shoes/attributedata/womens_high_heels/1/img_womens_high_heels_1041.jpg;data/shoes/attributedata/womens_high_heels/0/img_womens_high_heels_237.jpg;are shiny, not matte

data/shoes/attributedata/womens_rain_boots/1/img_womens_rain_boots_1016.jpg;data/shoes/attributedata/womens_boots/0/img_womens_boots_138.jpg;have no buckle or wedge heel
...

FIGURE 3. Example of reference image path and target image path with captions.

3.2. **Experiment details.** Our proposed framework for image retrieval system shown in Fig. 4.

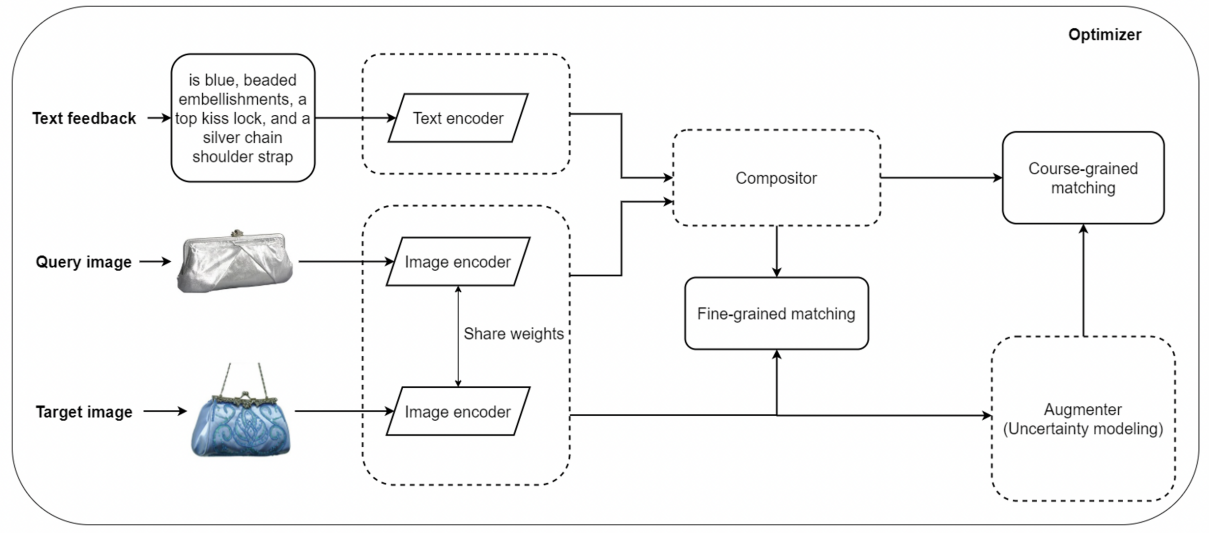

FIGURE 4. Image retrieval with text feedback pipeline via Multi-grained Uncertainty Regularization.

The following are the essential steps in the procedure for doing image retrieval with text feedback based on Multi-grained Uncertainty Regularization. The goal of this study is to use techniques that are thought to be superior to those used at each stage of this model. The study will use experimental data to show that the newly applied method is really more effective and optimal on the Multi-grained Uncertainty Regularization model.

Image encoder: There are other versions of ResNet that use the same basic idea but have various numbers of layers. The form that can operate with 50 neural network layers is referred known as Resnet50. Resnet-18, ResNet-50, ResNet-101, and more well-known CNN models are included in the Torchvision library. When applying those techniques to the model to encode the image, we will utilize ResNet-50 as a baseline to compare with ResNet-152 and ResNet-18.

Finding the best model for image encoding is the goal of comparing these models; for this experiment, first we did not take GPU RAM loss into consideration and instead concentrated on assessment. Given that ResNet-152 models for classification on ImageNet validation have the lowest top-1 and top-5 error rates among shallow ResNet models, we want to find out if it is really more effective than ResNet-50 and ResNet-18.

The image encoders are given by the output from layer 4 of the backbone networks. For ResNet-18, this layer will output a feature map with 512 channels, and for ResNet-50 and ResNet-152, there are 512 channels. ResNet-50 and ResNet-152's original final classification layer for ImageNet is removed. Because of this, the output feature from these two has 2048 dimensions, and we use a linear layer to convert the result to a 512-dim feature for additional comparison. We run these three models on the same config, which has 15 epochs and batch size is 32. The result will be visualized on Weights & Biases - a developer tool for Machine Learning. The ground truth recall status is shown by Recall@K. We present the best Recall@1, Recall@10, and Recall@50 of each query, following previous studies. The results are presented in Table 1.

Recall@10 rate is required for the uncertainty learning-based image retrieval with text feedback on the Shoes dataset in order to compare the accuracy. Though on this table, the Recall@10 result shows the better model is ResNet-50 (39.66), ResNet-152 also shows a

TABLE 1. Image encoders - Results on Shoes dataset.

	Recall@10	Recall@50	Recall@1	Average
ResNet-18	38.97	70.02	9.88	39.62
ResNet-50	**39.66**	70.62	9.34	39.87
ResNet-152	39.32	**73.60**	**9.97**	**40.96**

close result with 39.32 rates. However, when we observe through 15 epochs, the Recall@10 of ResNet-152 is higher than ResNet-50 for almost all epochs, which means the average Recall@10 of the mod-el when applying ResNet-152 performs better than ResNet-50. The average result also shows us that the image retrieval with text feedback model has been improved when we replace ResNet-50 with ResNet-152 as an image encoder.

The reason why ResNet-152 has more effective in this model can be explained by understanding why ResNet was born. ResNet uses residual blocks to increase the models' accuracy. The strength of this kind of neural network is the idea of "skip connections", which is at the foun-dation of the residual blocks. These skip connections operate in two different ways. First, they resolve the problem of the vanishing gradient by creating a different shortcut for the gradient to use. They also give the model the ability to learn an identity function. By doing this, it is made sure that the model's higher levels don't perform any worse than its lower layers.

In order to determine whether the model performs better than the small layers in ResNet-50 or ResNet-18, we can increase the number of layers in the neural network, for which we utilize ResNet-152. That can be understood easily that the ResNet-152 has the best performance compared to ResNet-50 and ResNet-18.

Regarding loss, simply a high loss number typically denotes inaccurate output from the model, whereas a low loss value denotes fewer errors in the model. The performance of ResNet-152 when applied to the model also showed in the line graph below. The line

FIGURE 5. The training loss of image-text retrieval when applied ResNet-152, ResNet-50, and ResNet-18.

of ResNet-152 has the lowest value, which is 48.43 compared to 50.02 for ResNet-50 and 51.72 for ResNet-18.

As a result, when this model with ResNet-152 is used in a real-time system and deployed online, it will take longer for the model to calculate and provide the result every time the database up-dates or when the user requests to obtain the product. ResNet-152 produces a greater level of accuracy (3% at R@50) when used as an image encoder in this model, but when applied to online environments, image encoders with fewer classes will produce better benefits. The re-sults of our studies on ResNet-152 suggest that using the ResNet model with more layers to this image retrieval with text feedback model will increase accuracy and also show the difference when compared to other ResNets. The result of the training loss of image-text retrieval when applied ResNet-152, ResNet-50, and ResNet-18 shown in Fig. 5.

Text encoder: In the context of natural language processing, an encoder is a type of neural network architecture that is used to convert raw text input into a fixed-length representation that can be used for downstream tasks such as text classification, sentiment analysis, or language modeling.

RoBERTa is a type of encoder that is based on the Transformer architecture and is pre-trained on large amounts of text data. By utilizing RoBERTa as the text encoder, the model can learn to extract high-level features and representations from the input text that are useful for the image retrieval task.

After the text input is encoded by RoBERTa, a linear layer is applied to reduce the feature channels' dimensions from 768 to 512. This step is often referred to as dimensionality reduction or feature compression and is commonly used to reduce the number of parameters in the model and improve its computational efficiency.

Additionally, this experiment involves testing the performance of other text encoders, such as LSTM and DeBERTa, on the image retrieval task. LSTM (Long Short-Term Memory) is a type of recurrent neural network that is commonly used for sequence modeling tasks, while DeBERTa is another type of transformer-based encoder that has been shown to achieve state-of-the-art performance on several NLP benchmarks.

The result of RoBERTa is significantly improved compared to LSTM in CoSMo model because before the Transformer model was created in 2017, LSTM was a substitute for RNN in terms of short-term memory restriction, but it still had some drawbacks. Since 2017 when Transformer was introduced, The NLP landscape has been completely transformed by Google's BERT and more current transformer-based approaches, which outperform the state-of-the-art on a number of tasks.

RoBERTa alters BERT's pre-training by removing the Next Sentence Prediction (NSP) task and adding dynamic masking, which causes the masked token to change throughout the training epochs. Additionally, it was shown that larger batch sizes were more beneficial for the training process. Since encoding and analyzing the captions that describe the target images in image retrieval with the text feedback model are crucial for the accuracy of this model, so we should try to apply new transformer methods that could enhance the effectiveness. DeBERTa is an example. Using two innovative methods, DeBERTa enhances the BERT and RoBERTa models. One innovative method that DeBERTa uses is the disentangled attention mechanism, where each word in the input sequence is represented using two vectors that convey its content and location, respectively. The attention weights are computed between words using disentangled matrices on their contents and relative positions, which helps the model to better capture the interactions between words and their positions in the sequence.

In addition, DeBERTa replaces the output softmax layer with an improved mask decoder to anticipate the masked tokens during model pre-training. This allows the model

to better predict the masked tokens and improve its ability to handle incomplete or partial input sequences.

By utilizing these innovative pre-training methods, DeBERTa has been shown to achieve state-of-the-art performance on several natural language processing benchmarks, including the GLUE benchmark and the SuperGLUE benchmark. These improvements in pre-training methods can also potentially enhance the effectiveness of image retrieval with text feedback models by better capturing the semantics and relationships between image captions and target images.

The implementation of RoBERTa and DeBERTa is similar following the introduction of the Hugging Face Community. RoBERTa can function as both a decoder and an encoder (using solely self-attention), in which case a layer of cross-attention is placed in between the layers of self-attention.

RoBERTa must be initialized with the is_decoder parameter and add_cross_attention set to True in order to be utilized in a Seq2Seq model; an encoder_hidden_states is then anticipated as an input to the forward pass. Meanwhile, DeBERTa can perform both encode and decode functions without any config attribute.

However, In the case of the Shoes dataset with relatively small numbers of images and short text descriptions, the RoBERTa model is able to process the input text efficiently, and the additional capabilities offered by DeBERTa's improved mask decoder may not be fully utilized. Moreover, the decode step is typically not necessary for the image encoder in the context of image retrieval with text feedback. Therefore, using DeBERTa in this specific application may not provide significant performance improvements over RoBERTa. However, in scenarios with larger amounts of text data or more complex text inputs, the strengths of DeBERTa may be more advantageous. The experimental results comparing these methods are presented in Table 2.

Optimizer: The optimizer is an essential component of the image retrieval with text feedback model that is responsible for updating the model parameters during training. The goal of the optimizer is to minimize the loss function, which is a measure of how well the model is performing on the training data.

There are several types of optimizers that can be used in the image retrieval with text feedback model, including stochastic gradient descent (SGD), Adam, and Adagrad. Each optimizer has its own advantages and disadvantages, and the choice of optimizer often depends on the specific requirements of the task and the characteristics of the data.

The optimizer works by computing the gradients of the loss function with respect to the model parameters, and then updating the parameters in the direction of the negative gradient. This process is repeated iteratively until the loss function reaches a minimum, at which point the model is considered to be trained.

With the optimizer algorithm in this model, we can barely deploy Stochastics Gradient Descent (SGD) with no momentum to find the minimum weight and bias with the lowest model loss. However, due to the fact that the cost function graph for deep learning is not convex, there are several problems with the SGD that cause it to perform less well than it should, including: 1) a propensity to converge to local minima rather than the global

Table 2. Text encoders - Results on Shoes dataset.

	Recall@10	Recall@50	Recall@1	Average
LSTM	36.23	68.13	**9.62**	37.99
RoBERTa	39.66	70.62	9.34	39.87
DeBERTa	**40.32**	**71.76**	9.25	**40.44**

minimum, 2) since the slope is changing gradually, the rate of change will also slow down, which will affect training, 3) the huge curvature, which was often high in non-convex optimization, will be challenging to traverse.

To address this issue, SGD with momentum was introduced, which uses an exponentially weighted moving average of the past gradients to smooth the updates and move more directly towards the optimum. It helps to accelerate convergence, especially in the presence of high curvature, noisy gradients or sparse gradients.

Based on the original study, we deploy the SGD with a momentum value is 0.9 and the learning rate is 2x10-2. The weights will be updated after each iteration by follow formula:

$$W_t + 1 = W_t - V_t \tag{1}$$

We need to calculate V_t so that it carries both the information of the slope (i.e. the derivative) and the information of the momentum, i.e. the previous velocity. V_t is determined as follows:

$$V_t = \ \beta * \ V_t - 1 + \ \eta \Delta W_t, with \ \beta \in [0, 1] \tag{2}$$

In this experiment, the momentum coefficient value 0.9 is β value because if $\beta = 0$, the optimizer will perform the same as SGD, and if $\beta = 1$, there is no decay. So the β value should be 0.9, 0.99, or 0.5 only.

While SGD with momentum helps to find the global minimum and improve the speed of SGD, it still has some drawbacks. Momentum still takes a while to stop when it gets closer to the end.

Another approach, known as Nesterov accelerated gradient (NAG). This is an extension of SGD with momentum that uses a modified update rule that first computes an approximation of the future position of the parameters before computing the gradient. This technique can lead to faster convergence than regular SGD with momentum. We calculate Vt with Nesteov momentum by using follow formula:

$$V_t = \ \beta * \ V_t - 1 + \ \eta \Delta W_t J(W_t - \beta * \ V_t - 1) \tag{3}$$

In Equation (3), our parameters' approximate future positions can be determined by computing $W_t - \beta * V_t - 1$, giving us a general idea of where they will be.

While SGD with Nesterov momentum is a powerful optimization algorithm that can improve the convergence speed and accuracy of the model, it still has some limitations. One of the main drawbacks of Nesterov momentum is that it can be sensitive to the choice of hyperparameters, such as the momentum coefficient and the learning rate. Moreover, Nesterov momentum requires computing the gradient of the objective function with respect to the future position of the parameters, which can be computationally expensive for complex models with a large number of parameters.

To address these limitations, the Adam (Adaptive Moment Estimation) optimizer was developed as a more advanced optimization algorithm that combines the advantages of both momentum techniques and adaptive learning rates [9]. Adam uses a different update rule that estimates the first and second moments of the gradient and adapts the learning rate for each parameter based on their gradients' past history. It also includes bias correction terms to correct for the initialization bias and scale of the first and second moments. However, one of the main drawbacks of Adam is that it can exhibit poor convergence behaviors for problems with sparse gradients or high dynamic range. RAdam addresses this by rectifying the adaptive learning rate parameter, which ensures that the learning rate doesn't become too large in the early stages of training [11]. It also includes a dynamic variance correction that reduces the variance of the adaptive learning rate estimator. Moreover, RAdam includes a warm-up mechanism that smoothly increases

TABLE 3. Optimizers result on Shoes dataset.

	Recall@10	Recall@50	Recall@1	Average
SGD with momentum	39.66	70.62	9.34	39.87
SGD with Nesterov momentum	39.98	72.74	9.68	40.80
RAdam	**41.87**	**74.14**	**10.08**	**42.03**

TABLE 4. Total training time and loss value of the optimizers.

Methods	Total training time	Loss
SGD with momentum	4h 11m 49s	50.02
SGD with Nesterov momentum	4h 2m 54s	49.88
RAdam	**3h 49m 53s**	**48.34**

the learning rate during the early stages of training to prevent unstable updates. This mechanism is particularly useful for large-scale models or problems with a large number of parameters.

For above reasons, we decided to do the experiment on RAdam. We apply the new learning rate equal to 2 x 10-4 to adapt with RAdam. The result after we deploy 3 optimizers is shown in Table 3. The results show that the more effective the optimizer, the higher the recall rate we get. We also consider training time and the loss value as important factors in comparing these three methods. The total training time and loss value for each of them are shown in Table 4.

As the optimizers that speed up the algorithm's convergence more effectively, the total training time is shorter and the model can get closer to the lower loss value. After comparing these optimizers, we can see that RAdam has the highest recall rate on Shoes dataset, as well as the total training time and loss, are the lowest.

In image retrieval with text feedback, the training speed can have a significant impact on the overall system's performance. Faster training times can allow for larger-scale experiments and can lead to better models, as more data can be used in the training process. In addition, faster training times can also help reduce the cost of model development and training by decreasing the time required to process large datasets. This can be particularly important in industries such as e-commerce or online advertising, where the speed of model deployment can impact busi-ness performance. However, it is important to note that faster training times can come at a cost, as some optimization methods or hardware configurations that enable faster training may also result in lower model accuracy. Therefore, it is important to carefully balance training speed with other factors such as model accuracy and interpretability when developing image retrieval systems with text feedback.

Overall, depending on the use case and application, the significance of training speed in image retrieval with text feedback will vary. Faster training times are still preferred in general, provided the model's performance is unaffected.

3.3. **Final results.** On the Shoes dataset, we experimented with applying certain methods that are supposed to be more optimal than those such as image encoding, text encoding, and optimizer algorithms based on the proposed model of Image retrieval with text feedback - Multi-grained Uncertainty Learning [4]. The outcomes found are demonstrated below:

- *For the image encoder*: we examined ResNet-18 and ResNet-152 in order to compare accuracy and loss results with ResNet-50. ResNet-152 demonstrates, in accordance with the experimental findings in the preceding chapter, that increasing the number of calculation layers increases the model's accuracy, but at the expense of resources used to train the model; in contrast, ResNet-18 and ResNet-50 use about the same resources. After testing three different ResNet models, it has been concluded that employing a ResNet model with fewer classes will speed up image retrieval system processing and have more tangible advantages if the accuracy difference is not too great (0.5-2%).
- *For text-encoder*: we have experimented with LSTM model and two transformer models, RoBERTa and DeBERTa. Due to its limitations before transformers were introduced, the LSTM model produced quite poor results, while the most widely used RoBERTa model produced results that were superior to those of the LSTM. Additionally, we can see that the model's accuracy is slightly enhanced (1.14%) after using DeBERTa on the Shoes dataset. Both the RoBERTa and DeBERTa models perform exceptionally well given the characteristics of the dataset, together with the length and context of the caption.
- *For the optimizer*: Despite its widespread use, SGD still has issues that make convergence take longer. We tried replacing the old optimizer with an SGD with Nesterov momentum and found that the results were slightly better. But after using Rectified Adam, the findings show that the model's accuracy and training speed have both increased (by around +4% with R@50). It is clear that, in addition to the accuracy, the issue of access speed is crucial for image retrieval with text feedback model when provided online. Based on actual evidence, we may recommend employing RAdam as a substitute for SGD with momentum in the prior study.

4. **Conclusions and future works.**

4.1. **Conclusions.**

Customers have enjoyed great shopping experiences for a very long time using text search (keywords), but as the number of products in the database grows, customer demand has changed diversely, making it more difficult for customers to find the products they want to buy by providing the product's features, therefore text search is no longer a good option. Image search has emerged as a promising solution to improve the user experience for product search, as it allows users to visually explore products and find what they are looking for more easily.However, image search still faces challenges, such as the issue of providing relevant and personalized search results to the user. One solution to this problem is the use of both text and images in the search engine. By combining these two modalities, users can provide more de-tailed and specific search queries that help the search engine better understand their preferences and needs.

Research on product retrieval models using image and text feedback is gaining popularity due to the way this model's features can enhance the customer experience. To improve the retrieval model, researchers are focusing on various aspects, including image feature extraction, text fea-ture extraction, image-text composition, and training algorithms. These advancements in the product retrieval model can lead to more accurate and personalized search results for the user, ultimately improving their experience.

In recent studies, researchers have used pre-trained models to optimize the extraction of image and text features. By doing so, they can improve the efficiency of the retrieval model and reduce the resource utilization and training time required. Moreover, novel techniques are being developed to address various challenges in product retrieval, such

as fine-grained attribute classification and cross-modal matching between image and text features.

While conducting this research, we realized that it might be challenging to interpret text input data and that it can be tough for a computer to figure out exactly what a user wants. In the ex-periments, when receiving input data, performing on only one representative dataset can some-times make it difficult to generalize and assess the accuracy of the outcomes, but this experimental dataset also shows how natural, situation-appropriate language is used.

We believe that, after analyzing the research results, improving the steps in the image retrieval with text feedback model will be the key to expanding the model's applicability. The training capacity and the results are going to keep improving as more datasets are created to further this concept. When the model is tested on the appropriate datasets, we can apply the idea of this model to other products on the e-commerce platform in addition to fashion data.

4.2. **Future works.** While we are pleased to have met the initial objective of our study, we recognize that there are still challenges and limitations in the field of image retrieval with text feedback. Therefore, we plan to broaden our research to address these issues and advance this field further.

- *Object detection in image feature extraction*: We cannot ensure that the user-provided image is properly suited for the image feature extraction model to deliver excellent results when it comes to object detection in images. Due to the image's abundance of confusing items, the Query Image may not be identified. As a result, the next step for the image feature extraction model will be to train it on a more realistic collection of data, then figure out how to make the object detection model more effective when used with the current model.
- *The interpretability of the text input data*: As we have noted, natural language processing is a complex task, and it can be challenging for computers to understand exactly what a user wants. To address this challenge, we plan to explore new techniques for text analysis and interpretation, such as incorporating machine learning algorithms that can better capture the nuances of language and user intent.
- *Datasets*: We must train the image retrieval with text feedback model on more datasets of greater size and difficulty. We need to utilize the challenge of extracting text features from datasets with nearly natural language captions in order to adapt to the varied user requirements, alongside training on simple datasets with attribute-like features.
- *Image-text compositor*: The image-text compositor is a crucial component of image retrieval with text feedback approach. We can do research to optimize this compositor or try to apply previously studied compositors in this study, to see if their results really work better than the old one or not. The input text and image data must also be taken into account when choosing the image-text compositor in order to make the appropriate adjustments.
- *Optimization of model training time*: In order to use the model on an e-commerce platform where data is frequently accessed and updated, our model must quickly adapt to these changes in order to provide the most accurate results. Every stage of the model's training process can be adjusted to reduce the training time while maintaining a high level of accuracy by exploring a variety of approaches.

Acknowledgment. This research is funded by University of Economics and Law, Vietnam National University Ho Chi Minh City / VNU-HCM.

REFERENCES

[1] Anwaar, M. U., Labintcev, E., and Kleinsteuber, M.. Compositional learning of image-text query for image retrieval. In *Proceedings of the IEEE/CVF Winter conference on Applications of Computer Vision*, 2021, pp. 1140-1149.

[2] Baldrati, A., Bertini, M., Uricchio, T., and Del Bimbo, A. Conditioned and composed image retrieval combining and partially fine-tuning CLIP-based features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4959-4968.

[3] Chen, Y., Gong, S., and Bazzani, L.. Image search with text feedback by visiolinguistic attention learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3001-3011.

[4] Chen, Y., Zheng, Z., Ji, W., Qu, L., and Chua, T. S.. Composed Image Retrieval with Text Feedback via Multi-grained Uncertainty Regularization. arXiv preprint arXiv:2211.07394, 2022.

[5] Devlin, J., Chang, M. W., Lee, K., and Toutanova, K.. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Dodds, E., Culpepper, J., and Srivastava, G.. Training and challenging models for text-guided fashion image retrieval. arXiv preprint arXiv:2204.11004, 2022.

[7] Hoffer, E., and Ailon, N.. Deep metric learning using triplet network. In *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015*, 2015, pp. 84-92.

[8] Isola, P., Lim, J. J., and Adelson, E. H.. Discovering states and transformations in image collections. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1383-1391.

[9] Kingma, D. P., and Ba, J.. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[10] Lee, S., Kim, D., and Han, B.. Cosmo: Content-style modulation for image retrieval with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 802-812.

[11] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., and Han, J.. On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265, 2019.

[12] O'Shea, K., and Nash, R.. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458, 2015.

[13] Pennington, J., Socher, R., and Manning, C. D.. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[14] Seymour, T., Frantsvog, D., and Kumar, S.. History of search engines. *International Journal of Management & Information Systems (IJMIS)*, 15(4), 2011, pp. 47-58.

[15] Saravanan, R., and Sujatha, P. (2018, June). A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In *2018 Second international conference on intelligent computing and control systems (ICICCS)*, 2018, pp. 945-949.

[16] Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L. J., Fei-Fei, L., and Hays, J.. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6439-6448.

[17] Zhang, Y., Xiang, T., Hospedales, T. M., and Lu, H.. Deep mutual learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4320-4328.