Contrastive Learning for Topic-Dependent Image Ranking

Jihyeong Ko[†], Jisu Jeong, Kyumgmin Kim

Abstract In e-commerce, users' feedback may vary depending on how the information they encounter is structured. Recently, ranking approaches based on deep learning successfully provided good content to users. In this line of work, we propose a novel method for *selecting the best from multiple images considering a topic*. For a given product, we can commonly imagine selecting the representative from several images describing the product to sell it with intuitive visual information. In this case, we should consider two factors: (1) *how attractive each image is to users* and (2) *how well each image fits the given product concept (i.e. topic)*. Even though it seems that existing ranking approaches can solve the problem, we experimentally observed that they do not consider the factor (2) correctly. In this paper, we propose CLIK (**C**ontrastive **L**earning for topic-dependent **I**mage ran**K**ing) that effectively solves the problem by considering both factors simultaneously. Our model performs two novel training tasks. At first, in *Topic Matching*, our model learns the semantic relationship between various images and topics based on contrastive learning. Secondly, in *Image Ranking*, our model ranks given images considering a given topic leveraging knowledge learned from *Topic Matching* using contrastive loss. Both training tasks are done simultaneously by integrated modules with shared weights. Our method showed significant offline evaluation results and had more positive feedback from users in online A/B testing compared to existing methods.

[†] Work done while intern at NAVER CLOVA.

Jihyeong Ko
WATCHA Inc., Seoul, South Korea. e-mail: louis.ko@watcha.com

Jisu Jeong
NAVER CLOVA, NAVER AI LAB, Seongnam, South Korea. e-mail: jisu.jeong@navercorp.com

Kyungmin Kim
NAVER CLOVA, NAVER AI LAB, Seongnam, South Korea. e-mail: kyungmin.kim.ml@navercorp.com

1 Introduction

In e-commerce, how information is composed for a product or an advertisement is essential to get positive user feedback. From infinite contents, it is important to give users information they want explicitly or implicitly. By ranking information, we can selectively provide information that satisfies users' tastes where the ranking method is actively covered in *Learning to Rank* [17, 3, 4, 2, 23]. Based on various *Learning to Rank* methods, many service platforms struggle to obtain positive user feedback by providing good content in web searches, product recommendations, or advertisements.

In particular, *creative ranking* is recently known as a ranking approach for product advertisement. It is for selecting a creative to advertise expected to attract users' attention when composing content for a product to sell. ByteDance, a technology company operating content platforms, has improved its advertising system based on creative ranking [34]. Their proposed PEAC (Pre Evaluation of Ad Creative model) is a pairwise ranking model using the various information of each candidate creative. By ranking creatives at the offline phase, it produces only potential creatives to their online recommender system. It has eventually improved the performance of their overall system. In addition, Alibaba, one of the largest online shopping malls, proposed another method using creative ranking [30]. They pointed out that PEAC is not flexible as it works only offline and then proposed a solution model VAM-HBM (Visual-Aware ranking Model and Hybrid Bandit Model) that works online. With VAM that captures visual information of creatives for ranking, they alleviated the cold start problem of typical methods based on a multi-armed bandit algorithm [27].

Fig. 1 Example of our problem: *selecting the best from multiple images considering a topic*. The main image should be selected considering (1) how attractive each image is to users and (2) how well each image fits a given topic.

Meanwhile, there is a problem similar to creative ranking in e-commerce: *selecting the best from multiple images considering a topic*. As in Figure 1, suppose that four products are on sale together within topic '*Cropped Sweatshirt for Women*'. In this case, many e-commerce platforms expose one of the product images as the main image to give users visual information about the topic. In Figure 1, the first image of cropped sweatshirt can be selected as the main. However, it is not easy to select the main in practice. It is hard to check whether each image matches the topic since there are often hundreds of products in a topic in real service. It is more cumbersome if there are off-topic products in the list (e.g. the fourth product in

Figure 1). In addition, we should be careful not to choose a low-quality product image, even if it is suitable for the given topic (e.g. the second product in Figure 1), as it can get less attention from users. The larger the number of products in a list, the more inefficiently time-consuming to perform the task with only human resources. Therefore, an automatic algorithm or a model is required for this problem.

Then, how do we select the best as the main image automatically? What should a model consider to pick the best? It may be suboptimal only to consider how each image catches users' attention (e.g. predicts user click-through-rate for each image) because there is no consideration of the relationship between given images and a topic. Even if the image of the pants for women in Figure 1 is attractive enough, it should not be selected as the main because it is totally out of the given topic. A model should understand the semantic relationship between given images and a topic to make a reasonable choice. As a result, to solve the problem, we have to deal with two factors:

(F1) *how attractive each image is to users*,
(F2) *how well each image fits a given topic*.

The problem may be solved by existing ranking methods that predict each image's ranking score based on representations of the images and topic. We can select the best by comparing the scores. However, we experimentally observed that they do not take the factor (F2) into account sufficiently. Despite using the representation of the topic and images, they cannot guarantee that the image of women pants will never be selected as the main in Figure 1. It is because they cannot penalize off-topic images appropriately. An additional method should be applied to overcome the limitation.

The factor (F2), consideration of the relationship between given images and topic, is closely related to the retrieval task. It is a task to search for data compatible with a query, and the search is often performed by measuring distances between embeddings. Recently, contrastive learning has shown successful performance at the retrieval task in various modalities [15, 11, 18]. Optimizing contrastive loss called InfoNCE [28] or also known as NT-Xent [5], a model minimizes distances between semantically similar embedding pairs (i.e. positive pairs) and maximizes distances between dissimilar pairs (i.e. negative pairs). Since the learning procedure is directly related to the retrieval task, contrastive learning is a key for the factor (F2). In summary, we leverage a ranking method for the factor (F1) and contrastive learning for the factor (F2). The most important thing here is to consider both factors *simultaneously*. In other words, if several images and a topic are given, a solution model should subordinate the semantic relationship between the topic and the images to its ranking scores.

In this paper, we propose a novel model CLIK (**C**ontrastive **L**earning for topic-dependent **I**mage ran**K**ing) to solve the problem of *selecting the best from multiple images considering a topic*. It can consider both factors above by performing two significant training tasks: *Topic Matching* and *Image Ranking*, as in Figure 2. At first, in *Topic Matching*, our model understands the semantic relationship between images and topics using contrastive learning inspired by CLIP [25]. Secondly, in *Image Ranking*, our model uses contrastive loss to rank given images considering

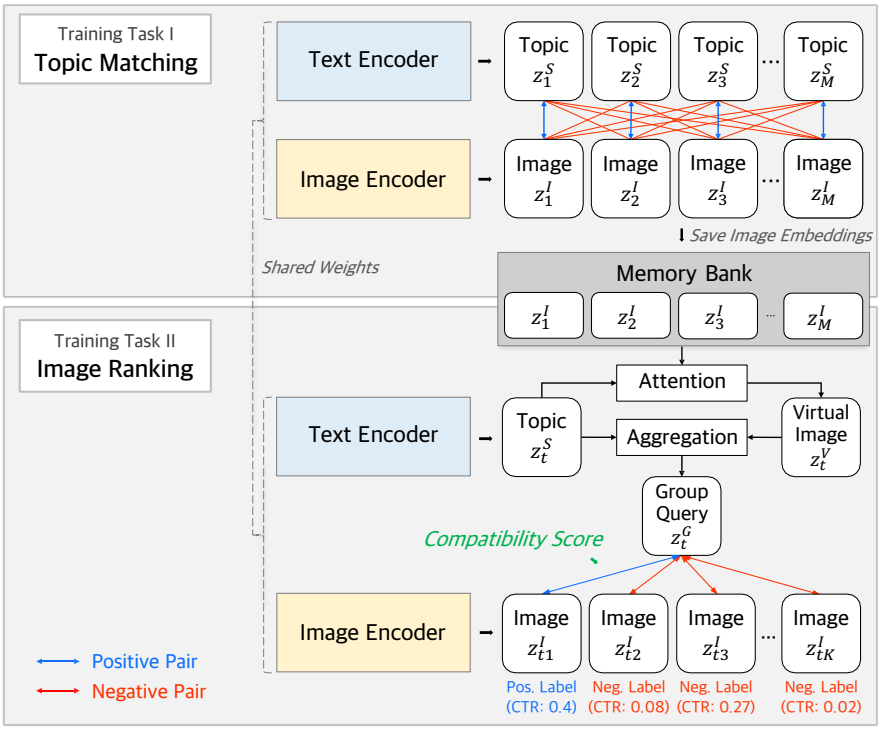

Fig. 2 Structure and training tasks of CLIK.

a given topic, leveraging knowledge learned from *Topic Matching*. Both tasks are done simultaneously by integrated modules with shared weights. As a result, our model successfully subordinates the semantic relationship between given images and a topic to ranking scores. From the offline experiments comparing our model with several baselines, we observed that our model shows significant performance and reasonably considers the semantic relationship between given images and a topic during ranking. We also applied our model to one of the services of our platform relevant to the main problem. From online A/B testing, we got more positive feedback from users compared to baseline.

2 Related Work

2.1 Creative Ranking

In the advertising system, *creative ranking* is a method to select a creative expected to attract users' attention when composing content for a product to sell. In general,

many creatives for a product are dynamically selected by online ranking algorithm. In this case, cold start problem can arise before each creative get enough impressions that are needed to make ranking result reliable. In order to treat this problem, many studies have been progressed. Bytedance proposed PEAC [34], a pairwise ranking model using representations of images, texts embedded in the images, and category information of creatives. By ranking creatives offline, it produced only reliable creatives to their own online recommender system rather than randomly sampled ones causing cold start problem. It eventually improved performance of their system. Alibaba proposed another creative ranking-based method [30]. They pointed out that PEAC of ByteDance is not flexible because it works only at the offline, and then proposed VAM-HBM receiving online observations flexibly, which consists of creative ranking method and MAB algorithm. In particular, with its creative ranking part VAM capturing visual information of creatives, they alleviated cold start problem of typical MAB-based methods.

Our problem is related to creative ranking in terms of selecting the best from a given list by ranking. However, there is a difference: *consideration of a topic*. For given images and topic, we experimentally observed that general creative ranking methods cannot explicitly consider semantic relationship between images and topic well. In order to inject understanding of it into a solution model, we should use another approach additionally, not only use creative ranking.

2.2 Contrastive Learning

Intuitively, *contrastive learning* (CL) is a learning method based on data representation comparison. The goal of it is to make similar pairs closer and dissimilar pairs far apart. Unlike general supervised learning requiring human annotation, CL only needs to define a similarity distribution obtained easily through pretext tasks (e.g. data augmentation) [14, 19]. It is, therefore, one of the popular learning approaches for self-supervised learning (SSL). In computer vision, various CL-based SSL methods[31, 5, 6, 12, 7] showed successful performances for many downstream tasks outperforming supervised learning-based pre-training (e.g. pre-training with image classification on ImageNet [8]), and recent studies to apply it to other modalities are also active [33, 26, 16].

In particular, CLIP [25] shows the significance of cross-modal CL pre-training with large-scale text-image pairs collected from websites. It learns multimodal embedding space of text and image. The authors demonstrated that CLIP performed competitively with state-of-the-art methods for many downstream tasks at computer vision and mentioned the potential for widely-applicable retrieval tasks on image, text, and image-text (cross-modal). Afterward, Jia et al. [15] proposed ALIGN using the same training procedure as CLIP but more data for training and proved that it outperformed performance in various retrieval tasks. Following the successful results, research on applying the broad utility of CLIP to the e-commerce field also proceeds [13].

We determined that CL would be appropriate for the factor (F2), that is, to understand the semantic relationship between images and topics, because CL has proven successful performance on retrieval tasks closely related to the factor (F2). Primarily, we adopted the training procedure of CLIP. Since the modality of the topic is often linguistic in the e-commerce field, training based on image-text comparison is an effective way to learn the semantic relationship between images and topics.

3 Method

In this section, we introduce our proposed CLIK. At first, we explain our problem, *selecting the best from multiple images considering a topic*. Second, we show an overview of CLIK. We describe the model structure and its two significant training tasks. At last, we explain how both tasks work in detail.

3.1 Problem Definition

The problem of *selecting the best from multiple images considering a topic* is solved by ranking given images. We assume a score exists that evaluates each image from the aspect of our problem and denote it as *a compatibility score* c. Additionally, we define $\{s, X\}$ as a group G where s is a given topic and $X = \{x_i\}_{i=1}^{|X|}$ is a given list of images.

$$G = \{s, X\}$$
$$X = \{x_i\}_{i=1}^{|X|}$$

where the number of elements of X is greater than 1. Here, we define the modality of a topic s as text because most topic information are represented as text in many e-commerce services. As a result, our goal is to find a model f that predicts a compatibility score of each image from a given group, and then the best image can eventually be selected as follows:

$$\{c_i\}_{i=1}^{|X|} = f(G)$$
$$c_* =_{c_i} \{c_i\}_{i=1}^{|X|}$$

where c_i is a compatibility score of corresponding x_i, and c_* is the compatibility score of the best image from given X. The argmax can surely be replaced as the argmin if a lower c indicates better compatibility.

For each image, the score is determined by two factors: (1) *how each image is attractive to users* and (2) *how each image fits a given topic*. For instance, suppose a

topic of 'men pants for a trip to the beach' and some product images are given. In this case, the best image at least should not only be appealing enough to get users' attention but also describe the topic well. In other words, a model should predict bad compatibility scores for an image of women pants that are not fit for the given topic or an ugly image that is not attractive enough. For the ideal prediction, a model needs a reasonable label or ground truth information that indicates each image's compatibility. The problem then can be solved by various approaches depending on the labeling strategy (e.g. classification, regression).

It seems that the problem can be solved by typical ranking approaches such as *Learning to Rank* or creative ranking. For example, we can consider a ranking model that uses a given topic as a query and predicts the ranking scores of a given images by measuring the distance between each image and the query. In this case, does the query represent '*topic*' semantically? Experimentally, we found that it does not. It is hard to guarantee that the model understands the semantic relationship between the given topic and images well. For instance, we observed that the model could not explicitly penalize images irrelevant to the given topic. One of the challenges is that model should subordinate the semantic relationship between images and topic to compatibility scores during prediction.

3.2 Overview

The structure of CLIK is composed of dual encoders and auxiliary modules. Dual encoders are feature extractors for images and topica (see 'Text Encoder' and 'Image Encoder' in Figure 2). Auxiliary modules include three parts: *Aggregation*, *Attention*, and *Memory Bank*. In a nutshell, they are used to generate a special query embedding for a compatibility score prediction, one of the essential components of CLIK.

Our model performs two novel training tasks. The first one is *Topic Matching* (*TM*). In *TM*, the model learns the semantic relationship between images and topics. The second one is *Image Ranking* (*IR*). In *IR*, the model predicts compatibility scores of given images considering a given topic. The best image then can eventually be selected by comparing the scores. Both tasks are done simultaneously by integrated modules with shared weights.

3.3 Topic Matching

In *Topic Matching* (*TM*), CLIK understands the semantic relationship between various images and topics. Inspired by CLIP [25], a significant cross-modal contrastive learning model, we adopted its training procedure. For given M pairs composed of image and corresponding topic, the model predicts M correct pairs from M^2 possible pairs that include $M \times (M - 1)$ incorrect pairs. Since the representations of images and topics are only needed, crowd-sourced labels are not required. Therefore, we can

use a large amount of data efficiently with no limitation of supervision. To optimize the following $L_{matching}$, which is the same as NT-Xent loss [5], CLIK maximizes the similarity of correct pairs and minimizes that of the others.

$$L_{matching} = (L_{S2I} + L_{I2S})\ /2$$

$$L_{S2I} = -\frac{1}{M}\sum_{m=1}^{M}\log\frac{\exp\left(sim(z_m^S, z_m^I)/\tau\right)}{\sum_{i=1}^{M}\exp\left(sim(z_m^S, z_i^I)/\tau\right)}$$

$$L_{I2S} = -\frac{1}{M}\sum_{m=1}^{M}\log\frac{\exp\left(sim(z_m^I, z_m^S)/\tau\right)}{\sum_{i=1}^{M}\exp\left(sim(z_m^I, z_i^S)/\tau\right)}$$

where z_m^I and z_m^S denote an image and topic embedding of the mth group G_m in mini-batch, τ is a temperature parameter, and $sim(\cdot,\cdot)$ is a cosine similarity function. The dimension of all embeddings above is the same.

To compose a mini-batch, we make M pairs of an image and corresponding topic from M groups. A positive pair is made by sampling an image from each group and pairing it with its corresponding topic, and the images and topics that do not match are all regarded as negative pairs (see the 'Topic Matching' part in Figure 2). Comparing various images and topics, CLIK (especially dual encoders) learns embedding space which reflects the semantic relationship between images and topics. Leveraging the space, in the other training task *Image Ranking*, our model then subordinates the semantic relationship between given images and topic to compatibility scores.

3.4 Image Ranking

In *Image Ranking* (*IR*), for a given group, CLIK predicts the compatibility score of each image considering the representation of the given images and a topic. Then we can select the best image by comparing the scores. Our model performs metric learning using contrastive loss over cosine similarity between given images and a special query embedding. Optimizing the following loss function $L_{ranking}$, the model makes compatible images closer to the query and the others farther away from it.

$$L_{ranking} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(sim(z_i^G, z_{i*}^I)/\tau\right)}{\sum_{k=1}^{K}\exp\left(sim(z_i^G, z_{ik}^I)/\tau\right)}$$

where z_i^G is a special query embedding of the ith group G_i of a mini-batch, z_{i*}^I is the best image embedding out of K included image embeddings $\{z_{ik}^I\}_{k=1}^K$ from X_i, and N is a mini-batch size. A mini-batch is composed by sampling K images from N groups. The dimension of all embeddings above is the same.

For each sampled group, we label the most relatively compatible image among K images as positive and the others as negative. Comparing cosine similarity between

images and the query z^G, CLIK classifies the most compatible image from given K images. We then regard the cosine similarity $sim(z^G, z^I_k)$ as a compatibility score c_k of x_k.

3.4.1 Group Query

Group Query z^G is a special query embedding representing the overall information of a group G. It is one of the essential elements of CLIK as it helps our model successfully perform two training tasks simultaneously.

Until we adopted *Group Query*, we considered using a given topic z^S as a query. In this case, however, the performance was closely the same as random ranking that just randomly shuffles the given image list. We guessed that the cause of this disaster lies in *the pairing contradiction problem* between two training tasks. Since both tasks depend on the same distance metric (cosine similarity) between embedding pairs, there is a risk of collision when the model performs both tasks simultaneously. The pair in *TM* consists of various images and corresponding topics (i.e. 'image↔topic'). In *IR*, for a given group, if we adopt an embedding of a given topic as a query, compatibility scores will be defined as the distance between image and topic, the same composition as *TM*.

Due to the sameness, a collision occurs. Since *TM* has an inter-group characteristic, pairs between images and a topic from the same group are tentatively labeled as positive. On the contrary, since *IR* has an intra-group characteristic, only one pair that includes the most compatible image is labeled as positive for a given group. This discrepancy prevents CLIK from learning appropriate solution space. For instance, there are many cases where an image pulls to its corresponding topic in *TM* (labeled as positive) but pushes away from it in *IR* (labeled as negative). For this reason, the key to using *Group Query* embedding is to overcome the pairing contradiction. We found that CLIK performs both training tasks successfully with *Group Query*, which means that CLIK eventually subordinates the semantic relationship between given images and a topic to compatibility scores, one of the challenges for our problem.

A *Group Query* embedding z^G is generated based on the auxiliary modules, aggregating a given topic z^S and another special embedding called *Virtual Image* embedding z^V (see the 'Image Ranking' part in Figure 2).

$$z^G = Aggregation(z^S, z^V)$$

where *Virtual Image* embedding z^V is an embedding of a virtual image that semantically fits a given topic. An attention mechanism generates it. For a group G, the attention operation is performed by using the given topic z^S as a query and '*Memory Bank*', one of the auxiliary modules, as both keys and values as follows:

$$z^V = \sum_j \alpha_j z_j^I, \quad z_j^I \in \textit{Memory Bank}$$

$$\{\alpha_j\}_{j=1}^M = Softmax\left(z^S \odot \textit{Memory Bank}\right)$$

$$\textit{Memory Bank} = \left\{z_j^I\right\}_{j=1}^M$$

where *Memory Bank* stores memories of various images that CLIK has encountered. In *TM*, the model meets numerous images sampled from many groups. The image embeddings from *TM* are stored explicitly in the *Memory Bank*, and then are used to generate *Group Query* embedding. As model parameters are updated, we update the bank with newly extracted image embeddings for every training step to prevent the problem of stored embeddings being outdated [12]. We reported the ablation study for using *Group Query* embedding in detail in the experiment section.

3.5 Summary

To solve the main problem, CLIK performs two training tasks simultaneously: *Topic Matching* and *Image Ranking*. In *Topic Matching*, the model understands the semantic relationship between various images and topics. Optimizing $L_{matching}$, the model learns to determine which image matches which topic in semantic aspect. In *Image Ranking*, optimizing $L_{ranking}$, the model selects the most relatively compatible image from a given group by predicting a compatibility score for each image. Unlike typical ranking methods, CLIK predicts the scores considering the representation of given images and the semantic relationship between the images and the topic by leveraging knowledge learned from *Topic Matching*. As a result, CLIK minimizes loss function L_{CLIK} as follows:

$$L_{\text{CLIK}} = L_{matching} + \lambda \cdot L_{ranking}$$

where λ is a scalar to adjust the contribution of two loss functions. We set it to 20.

4 Experiments

In this section, we conduct experiments on a real-world dataset to evaluate CLIK. We reported offline and online result based on *Online Special Exhibition*, one of our services. At first, we explain *Online Special Exhibition* and how we collected data from the service. Secondly, we show offline evaluation results. Lastly, we show online evaluation results.

4.1 Online Special Exhibition

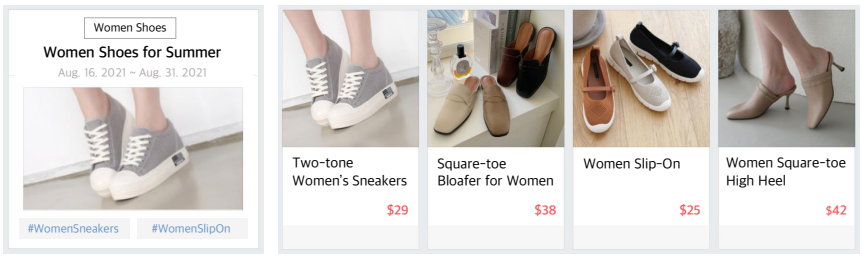

Fig. 3 Example of *Online Special Exhibition*. The first part is overall information of an exhibition, and the others are products of the exhibition. For CLIK, text embedded in the exhibition is used as *a topic*, product images are used as *a list of images*, and the exhibition is regarded as *a group*.

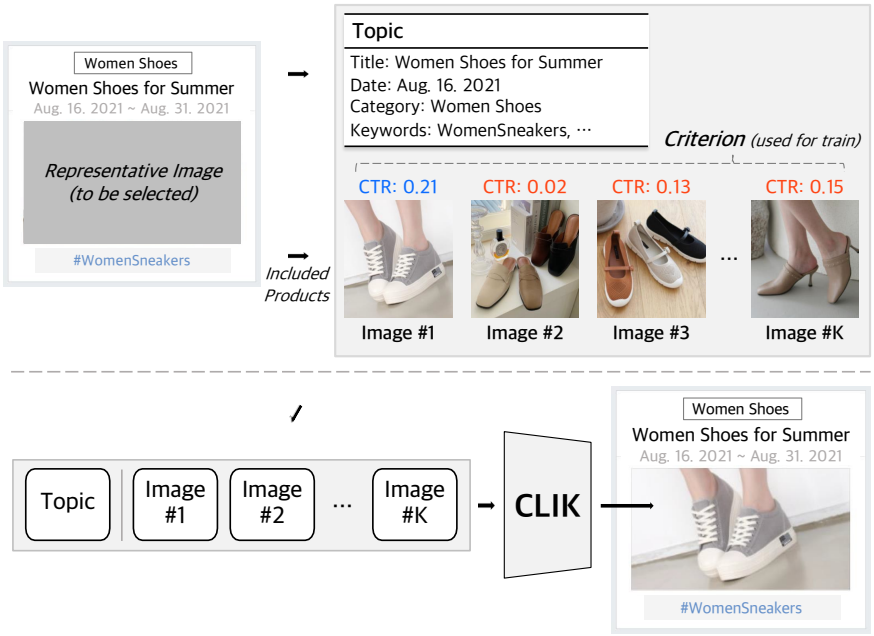

Fig. 4 *Representation Image Selection* by CLIK in *Online Special Exhibition*

Online Special Exhibition is a service that collects and sells products suitable for a special theme. On the main page, users can grasp at a glance the theme as in Figure 3. Each theme is described through not only text information such as title and

Table 1 Statistics of data collected from *Online Special Exhibition* service.

Type	Criterion	# Exhibitions	# Products	Date
1	CTR	1,605	104,716	Aug. 2021-Nov. 2021
2	Review Count	4,174	293,501	Aug. 2021-Nov. 2021

category but also a representative image. Especially, the representative image has been determined by our service operators recently. Since there often exist products that are off-themed from the corresponding exhibition or have low image quality, the operators should filter them delicately. This human-based process has problems in that personal tastes are subordinated to the selection, and it is inefficient to pick the representatives for hundreds of exhibitions daily. To overcome the problems, we tried to apply CLIK to the so-called '*Representative Image Selection*.' With CLIK, we can make automatic selections based on the estimated potential of each image as a representative image.

Figure 4 is an explanation for the application of CLIK. We regard an exhibition as *a group* G, the product images as *a list of images* X, and the text describing its theme as *a topic* s. In addition, since we collect various implicit user feedback on each product, we adopt one of them as an indicator or criterion for labeling to guide CLIK infer *a compatibility score* c for each image. For example, each product's user click-through rate (CTR) can be adopted as the criterion. In this case, for a given exhibition, the image of a product whose CTR is the greatest among the given products is labeled as positive, and the others are labeled as negative in IR. Note that user feedback generated from a service relevant to our problem can only be adopted as the criterion. For example, the user CTR of products from service with no topical information covering some products cannot be adopted as a criterion because the service is far from our problem.

In offline evaluation, we evaluated how accurately CLIK predicts the representative image based on two datasets. We adopted two metrics suitable for *Representative Image Selection*. Additionally, we observed how user feedback changes by applying CLIK online. As a result, we obtained successful results in both evaluations.

4.2 Data Collection

We collected *Online Special Exhibition* dataset from August to November 2021. For labeling, we conducted the collection process based on two labeling criteria: CTR and Review Count (i.e. the number of user reviews) for each product generated from *Online Special Exhibition*. Consequently, we collected two types of datasets.

Since some products are banned from sale or deleted, we dropped the exhibitions where less than 50% of the registered products are collected. For the CTR criterion

dataset (Type 1 at Table 1), we removed exhibitions with no clicks and just collected products whose impressions are 10 or more to use only products with reliable CTR. Additionally, we dropped the exhibitions where the uniqueness of CTR is less than 2 and the number of zero-CTR products is more than 5. On the one hand, for the dataset using the Review Count criterion (Type 2 at Table 1), we dropped exhibitions with no reviews and only collected products whose reviews were 10 or more. Also, we excluded the exhibitions where the uniqueness of review counts is less than 2 and the number of zero-review products is more than 5. Since the image sizes of products varied (e.g. 600×600, 1000×1000), we pre-resized all images for training efficiency. We collected title, keywords, publication date, and category information for topic information. Table 1 describes statistics of collected data. After the collection, we used 90% of them as train dataset and the other 10% as test dataset by random sampling.

4.3 Experiment Settings

4.3.1 Metrics

Since the main goal of our problem is to select the best image, we adopted mean reciprocal rank (*MRR*) and newly defined TopK-Top1 accuracy (*TopK-Top1*) as evaluation metrics to focus on the first-ranked image.

$$MRR = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{Rank(x_{n*}, f(G_n))}$$

$$TopK\text{-}Top1 = \frac{1}{N}\sum_{n=1}^{N}1(f(G_n), K)$$

where $Rank(x_i, f(G))$ is a predicted rank of image x_i by a model f for a given group G, and $1(f(G, K))$ is an indicator function that outputs 1 if the true rank of the image predicted as the best is less than K and 0 for the other cases. In short, *MRR* is for observing how the model predicts the actual first-ranked image, and *TopK-Top1* is for observing the actual rank of the image predicted as the best. The best and worst value of both metrics is 1 and 0.

4.3.2 Implementation Details

We adopt BERT [9] to encode topics and ViT [10] to encode images. They are known as scalable and efficient models using Transformer architecture [29]. We used BERT composed of 6 layers with 768 hidden dimension, ViT composed of 12 layers with 384 hidden dimension, and initialized them with pre-trained weights. The dimension of output embeddings from both encoders was set to 128. We adopted

a dot product-based attention module for *Attention*, a one-layered fully connected layer as *Aggregation* that maps concatenated embeddings of a topic embedding and *Virtual Image* embedding into a *Group Query* embedding of dimension 128. We added a tensor buffer for *Memory Bank* into our model by referencing MoCo [12].

We apply only a random crop to the input image with a size of 224×224 for the train and only resize it to the same size for the evaluation. We pre-processed text information of topics in '[CLS] *Title* [SEP] *Publication Date* [SEP] *Category* [SEP] *Keywords* [SEP]', the compatible input format for BERT. The maximum length of the text is defined as 128 with filling empty spaces with '[PAD]' tokens.

In *TM*, we set the batch size to 512 ($M = 512$), and the size of *Memory Bank* is the same. In *IR*, we set the number of sampled groups to 12 ($N = 12$), and each group consists of randomly sampled 20 images ($K = 20$) for the train. On the one hand, we set N to 1 and K to 50 for evaluation in *IR*, the numbers that better reflect our real online service. We use AdamW [21] applying weighted decay regularization to all weights with a decay rate of 0.1. We update the topic encoder (BERT) with an initial learning rate of 0.00005 and 0.0001 for the other parts and incorporate learning rate warm-up over the first 1% steps, followed by cosine decay [20] to lower the learning rate to zero. The temperature parameter τ is fixed as 0.07. We build our model and experiment settings based on PyTorch [24], a popular deep learning framework, and use the automatic mixed-precision [22] to accelerate training and save memory. We trained our model for 10 epochs with 4 P40 GPUs.

4.3.3 Baselines

To verify the significance of CLIK, we compare it with a few loss functions: *Triplet*, *Pairwise*, and *Pointwise* loss. They are general loss functions for *Learning to Rank* or creative ranking. We compared performance between the baselines and our model. Additionally, we observed the inference results to evaluate whether each model considers the semantic relationship between images and a given topic well for ranking.

- ***Triplet Loss*** With triplet loss, a model takes an anchor, a positive, and a negative. Then the model makes the positive closer to the anchor and the negative farther away than a margin. For a given group G, we use a topic s as an anchor and assign positive and negative to two randomly sampled images from X by comparing their values of the pre-defined criterion (e.g. CTR).

$$L_{triplet} = max\left(\|z^S - z^I_{pos}\|^2 - \|z^S - z^I_{neg}\|^2 + \alpha, 0\right)$$

  where z^S is an embedding of a given topic, z^I_{pos} and z^I_{neg} is positive and negative product image sampled from group G, and α is the margin set to 0.2. We then regard the distance between embeddings of a topic s and an image x as a compatibility score c of x.
- ***Pairwise Loss*** Pairwise loss is optimized by comparing a pair of samples as in [34, 2]. For our problem, we randomly sample two images from a group G at

first. Then, extract embeddings of each image and given topic s by corresponding encoders and concatenate each image embedding with the topic embedding. By forwarding both concatenated embeddings to a one-layered fully connected layer, model predicts score for each sampled image. The score then used for comparison for pairwise loss, and we regard it as a compatibility score c.

- ***Pairwise Loss*** A model with Pairwise loss optimizes the loss by comparing a pair of samples as in [34, 2]. For our problem, we randomly sample two images from group G. Then, we extract embeddings of each image and given topic s by corresponding encoders and concatenate each image embedding with the topic embedding. By forwarding both concatenated embeddings to a one-layered fully connected layer, the model predicts a score for each sampled image. The score is then used for comparison during optimization, and we regard it as a compatibility score c.

$$L_{pairwise} = -\left(y \log \sigma(c_i - c_j) + (1 - y) \log(1 - \sigma(c_i - c_j)\right)$$

  where c_i is a compatibility score of x_i, σ is a sigmoid function, and y is 1 if the value of the criterion of image x_i is greater than that of image x_j and 0 for the other case.
- ***Pointwise Loss*** To optimize pointwise loss, a model predicts scores of samples one by one. It is generally optimized by minimizing mean squared error between labels and predicted scores. Due to the hardness of actual value prediction, this approach is known to have lower performance than the pairwise loss that considers the only relative relationship of a pair [34]. We extract embeddings of each image and its corresponding topic and concatenate them. Then predict a score by forwarding the embedding through a one-layered fully connected layer. We define the criterion value for each product image as a label and regard the predicted score as compatibility score c.

$$L_{pointwise} = \frac{1}{N} \sum_{n=1}^{N} \left(y_{nj} - c_{nj}\right)^2$$

  where y_{nj} and c_{nj} are a value of criterion and a compatibility score for x_{nj} from group G_n.

For the baselines above, we set the same dual encoders and the same dimension of embeddings as CLIK for fairness (i.e. BERT and ViT for dual encoders and 128 for encoded embedding dimension).

4.4 Offline Evaluation

4.4.1 Comparison with baselines

Table 2 Offline evaluation compared to baselines.

	CTR				Review Count			
	MRR	Top1-Top1	Top3-Top1	Top5-Top1	MRR	Top1-Top1	Top3-Top1	Top5-Top1
CLIK	**0.1226**	**0.0496**	0.0729	0.1283	0.1627	**0.0828**	**0.1379**	**0.2103**
Triplet	0.102	0.0233	0.0758	0.1254	**0.1645**	0.0448	0.1000	0.1414
Pairwise	0.1063	0.0379	0.0641	0.1195	0.1380	0.0448	0.0828	0.1310
Pointwise	0.121	0.0379	**0.0947**	**0.1457**	0.1078	0.0207	0.0517	0.0724
Random	0.0899	0.02	0.06	0.1	0.0899	0.02	0.06	0.1

We compared CLIK with baselines using two types of datasets where one uses CTR and the other uses Review Count as a labeling criterion. According to the Table 2, CLIK shows significant performance overall compared to the baselines. We could conclude that CLIK is an especially suitable method for the *Representative Image Selection*. The first-ranked one is more important than the others because *Top1-Top1* accuracy is superior to the others. In addition, since the overall *TopK-Top1* accuracy is relatively high, CLIK is likely to predict at least a high-ranked image as the first more stably, even if it is not a first-ranked image.

Fig. 5 Inference result comparison between CLIK and the baseline with triplet loss. Other baselines show similar inference patterns to those of triplet loss.

In addition, we can see the vital characteristic of CLIK from the inference result comparison. The Figure 5 shows ranking results for a given product list of an exhibition, including 50 products. From the text, we can guess that the given exhibition's theme is relevant to pants for men (e.g. 'Title: Men Trousers, Bending, Spandex Pants'). According to the topic, the representative should visually depict pants for men. Therefore, the model should not rank images of given top products high, or the compatibility scores of images for the top products should be lower than those of bottom products. From this aspect, CLIK does its job much better than the baselines. According to the best-ranked 10 images in Figure 5, there are no top product images from CLIK, while the baseline includes several top products. Additionally, since the worst-ranked 10 images from CLIK are mainly composed of top product images, we can conclude that our model effectively subordinates the semantic relationship between given images and topic to the compatibility scores. On the other hand, since the inference result of the baseline shows a randomly mixed top and pants in worst-ranked and best-ranked 10 images, it seems that general ranking methods cannot capture the semantic relationship.

4.4.2 Usage of Group Query

Table 3 Experiment for query usage (Criterion: Review Count)

Query	MRR	Top1-Top1	Top3-Top1	Top5-Top1
Group Query	**0.1627**	**0.0828**	**0.1379**	**0.2103**
Virtual Image	0.1232	0.0483	0.1103	0.1655
Topic	0.0909	0.0207	0.0621	0.1000
Random	0.0899	0.02	0.06	0.1

Group Query embedding is one of the essential elements of CLIK. It helps the model avoid *the pairing contradiction problem* between two training tasks. For a given group, we could consider using a given topic as a query ('Topic' in Table 3) until we adopted *Group Query*. In this case, however, the performance was similar to random ranking, which randomly shuffles the given image list ('Random' in Table 3). On the other hand, according to the superior result of *Group Query* ('Group Query' in Table 3) compared to the case of *Topic*, we conclude that it is a key to overcoming the contradiction problem. With the new modality of *Group Query* combining images and topic, each pair composition of both tasks becomes different, and the model can eventually perform both tasks successfully simultaneously. Meanwhile, we tested an additional hypothesis. We expected that the *Virtual Image* embedding for generating *Group Query* could also prevent the contradiction problem in the same way as in the case of *Group Query*. From the result of *Virtual Image* in Table 3, we found that the performance is superior to the case of *Topic* even in this case. However, it is worse

than the case of *Group Query* where we can conclude that the combined information inherent in *Group Query* is helpful for CLIK.

4.5 Online Evaluation

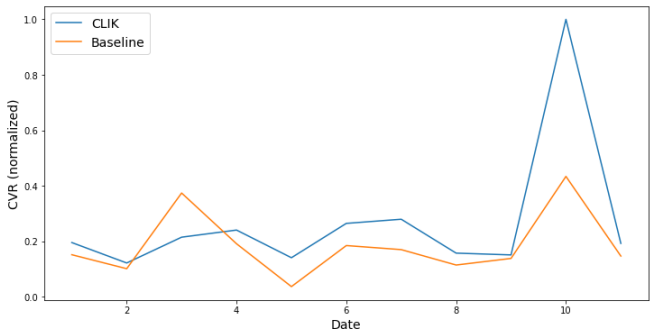

Fig. 6 Online A/B testing result. In the case of baseline, representative image for each exhibition is randomly selected from its product list. It shows an overall improvement of about 44% for user CVR when we apply CLIK to our live service than the case of baseline.

We analyzed the effect of CLIK in the real world through online A/B testing at our service *Online Special Exhibition*. Currently, in our service, *Representative Image Selection* is just made by randomly selecting one of the product images of each exhibition as the representative to use no human resources. Therefore, for the test, we compared CLIK with all our service users' random selection process tracking conversion rate (CVR). We conducted the test for 11 days.

The result is in Figure 6, where 'Baseline' is the original random selection process. We measured users' CVR with min-max normalization. On all days except one day during the test, normalized CVR was higher in applying CLIK than in the other case. In particular, we saw an overall improvement of about 44%, demonstrating CLIK's ability to produce good content for users with the sophisticated consideration required for *Online Special Exhibition*. We are actively applying CLIK to our service based on the successful results.

5 Conclusion

In this paper, we proposed CLIK for the problem of *selecting the best from multiple images considering a topic*. With two training tasks, our model solves the problem by understanding how each image is attractive to users and how each image fits a

given topic. We demonstrated that CLIK is superior to existing ranking methods and encourages positive feedback in our live service.

Despite the significance, our work has a limitation. We use one of the values of each image as a labeling indicator for ranking. Still, it is hard to guarantee that the value is determined only by the image's appearance or the relationship between the image and a given topic. For instance, when we adopt the CTR of each product as its value, CTR may be high not because the product image is attractive but only because of a special event at the service. Thus, reasonable refinement is required to train the model ideally. That is, we must rule out factors outside the assumptions of our problem to find a reliable value. Meanwhile, we further expect CLIK to be compatible with various modalities. Inspired by CLIP [25], CLIK only deals with texts and images. Since many contrastive learning approaches using various modalities have been proposed recently [1, 26, 32], we believe we can improve CLIK to perceive various modalities in the future. The modality-agnostic model will be of greater help in various e-commerce services.

In the e-commerce field, although we frequently face situations similar to *selecting the best from multiple images considering a topic*, solutions for them have not been actively studied. It is not easy to solve them optimally just with a general ranking approach because it does not consider many factors of the situations. We hope this work motivates future research to tackle these problems in various research groups or e-commerce platforms.

References

1. Akbari, H., Yuan, L., Qian, R., Chuang, W.-H., Chang, S.-F., Cui, Y., and Gong, B. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems 34* (2021), 24206–24221.
2. Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning* (2005), pp. 89–96.
3. Burges, C. J. From ranknet to lambdarank to lambdamart: An overview. *Learning 11*, 23-581 (2010), 81.
4. Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning* (2007), pp. 129–136.
5. Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (2020), PMLR, pp. 1597–1607.
6. Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297* (2020).
7. Chen, X., Xie, S., and He, K. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 9640–9649.
8. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (2009), Ieee, pp. 248–255.
9. 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).

10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
11. Gao, T., Yao, X., and Chen, D. Simsce: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).
12. He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), pp. 9729–9738.
13. Hendriksen, M., Bleeker, M., Vakulenko, S., Noord, N. v., Kuiper, E., and Rijke, M. d. Extending clip for category-to-image retrieval in e-commerce. In *European Conference on Information Retrieval* (2022), Springer, pp. 289–303.
14. Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., and Makedon, F. A survey on contrastive self-supervised learning. *Technologies 9*, 1 (2020), 2.
15. Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning* (2021), PMLR, pp. 4904–4916.
16. Jiang, D., Li, W., Cao, M., Zou, W., and Li, X. Speech simclr: Combining contrastive and reconstruction objective for self-supervised speech representation learning. *arXiv preprint arXiv:2010.13991* (2020).
17. Karmaker Santu, S. K., Sondhi, P., and Zhai, C. On application of learning to rank for e-commerce search. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval* (2017), pp. 475–484.
18. Krishna, T., McGuinness, K., and O'Connor, N. Evaluating contrastive models for instance-based image retrieval. In *Proceedings of the 2021 International Conference on Multimedia Retrieval* (2021), pp. 471–475.
19. Le-Khac, P. H., Healy, G., and Smeaton, A. F. Contrastive representation learning: A framework and review. *IEEE Access 8* (2020), 193907–193934.
20. Loshchilov, I., and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).
21. Loshchilov, I., and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
22. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. Mixed precision training. *arXiv preprint arXiv:1710.03740* (2017).
23. Mishra, S., Verma, M., Zhou, Y., Thadani, K., and Wang, W. Learning to create better ads: Generation and ranking approaches for ad creative refinement. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (2020), pp. 2653–2660.
24. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems 32* (2019).
25. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (2021), PMLR, pp. 8748–8763.
26. Shin, K., Kwak, H., Kim, S. Y., Ramstrom, M. N., Jeong, J., Ha, J.-W., and Kim, K.-M. Scaling law for recommendation models: Towards general-purpose user representations. *arXiv preprint arXiv:2111.11294* (2021).
27. Slivkins, A. Introduction to multi-armed bandits. *arXiv preprint arXiv:1904.07272* (2019).
28. Van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv e-prints* (2018), arXiv–1807.
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems 30* (2017).

30. Wang, S., Liu, Q., Ge, T., Lian, D., and Zhang, Z. A hybrid bandit model with visual priors for creative ranking in display advertising. In *Proceedings of the Web Conference 2021* (2021), pp. 2324–2334.
31. Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 3733–3742.
32. Ye, R., Wang, M., and Li, L. Cross-modal contrastive learning for speech translation. *arXiv preprint arXiv:2205.02444* (2022).
33. Yue, Z., Wang, Y., Duan, J., Yang, T., Huang, C., Tong, Y., and Xu, B. Ts2vec: Towards universal representation of time series. *arXiv preprint arXiv:2106.10466* (2021).
34. Zhao, Z., Li, L., Zhang, B., Wang, M., Jiang, Y., Xu, L., Wang, F., and Ma, W. What you look matters? offline evaluation of advertising creatives for cold-start problem. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (2019), pp. 2605–2613.