Automatic Data Generation of Incorrect Image-Text Pairs for Effective Contrastive Learning of CLIP Model

Rina Tagami
Chukyo University
Yagoto Honmachi
Showa-ku
101-2, Nagoya,
Japan
tagami@isl.sist.chukyo-u.ac.jp

Hiroki Kobayashi
Chukyo University
Yagoto Honmachi
Showa-ku
101-2, Nagoya, Japan
kobayashi@isl.sist.chukyo-u.ac.jp

Shuichi Akizuki
Chukyo University
Yagoto Honmachi
Showa-ku
101-2, Nagoya, Japan
s-akizuki@sist.chukyo-u.ac.jp

Manabu Hashimoto
Chukyo University
Yagoto Honmachi
Showa-ku
101-2, Nagoya, Japan
mana@isl.sist.chukyo-u.ac.jp

ABSTRACT

In this study, we proposed a method for automatically generating high-quality CLIP(Contrastive Language Image Pre-training) training data to improve the performance of text-based image retrieval using CLIP. In general, two types of image-text pair data are used in CLIP training: correct pairs and incorrect pairs. correct pairs are pairs in which the image and text content are compatible, and are created by scraping or other methods. incorrect pairs are incompatible image-text pairs, which are created by changing the combination of the correct pairs. CLIP is completed by contrastive training to increase the similarity between the image and text in correct pairs and decrease the similarity in incorrect pairs. However, when there are multiple images in the training data that are similar to each other, the text attached to them is also considered to be similar to each other, and although it is preferable to treat them as correct pairs, changed pairs are treated as incorrect pairs. In other words, incorrect pairs with high relevance between image texts are learned as having low relevance between image texts, and this inconsistency has a negative impact on the CLIP model. Therefore, if two images taken from the training data are not similar, then the similarity between texts assigned to them should also be low, so that a highly reliable incorrect pair can be created by exchanging the assigned text with each other. We applied this idea to the results of clustering the images and texts in the training data, respectively, and used the similarity between the clusters to generate an incorrect pair, then learned to increase the negative effect as the similarity between images was lower. The results of an experiment using the Amazon review dataset, which is commonly used in this field, showed a 21.0% improvement in Rank@1 score compared to vanilla CLIP.

Keywords

Large language Models, Image Retrieval, Image-Text Dataset, CLIP, Contrastive Learning, K-means Clustering

1 INTRODUCTION

The use of online shopping has increased in recent years due to the ease and convenience of purchase. Online shopping applications are equipped with a function that allows users to search for desired products by entering keywords or sentences into a search system. As the usage rate of the search system increases in proportion to the usage rate of the application, there is a need to improve the search accuracy. However, when keywords or text are input, images unintended by the user are sometimes output as search results. This study proposes a methodology for a search system that outputs images when text is input to output images that meet the user's intention.

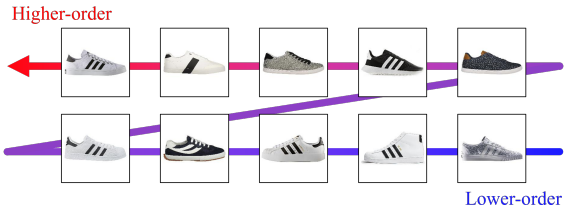

Figure 1: Output results when the text "Black and white striped sneaker." is entered into the vanilla CLIP (top 10).

Various image retrieval methods have been proposed, the most widely used of which [SIG08] assign keywords or phrases related to images as "tags" match the tags with text queries, and output images contain-

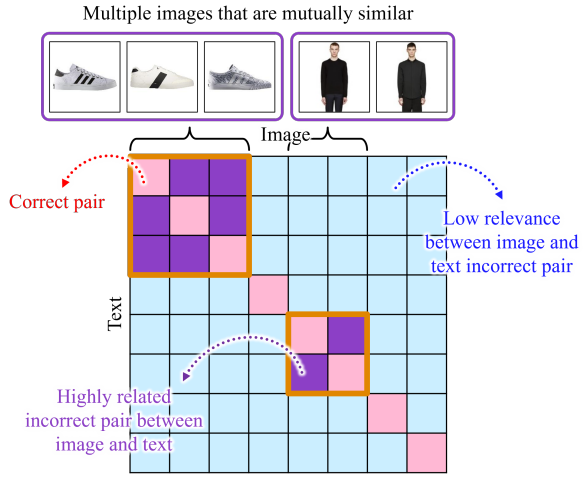

Figure 2: Relationship between correct and incorrect pairs.

ing tags with high similarity in a ranking format. The risk here that if the annotator tags an image incorrectly, or if the tagging is highly ambiguous, the result will be unintended. There are also automatic captioning techniques [KIL16] using deep learning and image retrieval [GIA15] based on the relationship between images and hashtags, but problems remain in terms of the cost of labeling training data and the reduced retrieval accuracy for unknown data. Recently, multimodal approaches [QI20][LI23] have become more available, and in particular, image retrieval using Contrastive Language Image Pre-training (CLIP) [RAD21] is becoming established [BAL22][HEN22]. Because of its zero-shot learning capability, CLIP can output relevant images even for text that is not in the training data. However, the problem of outputting images with low relevance to the input text, as shown in the Figure1, remains unsolved. A previous study [AGA21] on this issue suggested that CLIP performs well on image retrieval for general categories, but may perform poorly on certain tasks due to inherited biases. Other previous study [SHA23] suggests that the linguistic representation of images in a particular category or text describing that category is not well learned if what the image represents does not match the text prompt.

In the proposed method, the CLIP model is improved by modifying the current training data and adjusting the number of data to improve the retrieval accuracy. First, CLIP performs contrastive training using correct pairs (diagonal components in Figure2) and incorrect pairs (off-diagonal components in Figure2), which have high relevance between image and text content. When the training data consists of K pairs of images and texts, we assume that “all of them are correct pairs" and treat all the $K^2 - K$ generated by changing the combination quite of K pairs as incorrect pairs. This is reasonable when the pairs are independent, i.e., the images are not similar to each other. However, when there are multiple images in the training data that are similar to each other, the text attached to the images is also considered to be similar to each other, so all pairs are treated as incorrect pairs, even though it is preferable to treat the pairs with different combinations as correct pairs. Therefore, incorrect pairs (purple area in Figure2), whose contents are highly related to each other, are learned to be less related, which has a negative impact on the CLIP model. In addition, since there is a large difference in the number of data between correct and incorrect pairs, the learning is biased toward a large number of data.

In this study, we propose a method to solve both problems simultaneously. The main idea is to carefully select only the incorrect pairs that are expected to have low relevance between image-text pairs (blue region in Figure2), and adjust the number of data so that the number of incorrect pairs is the same as the number of correct pairs. Specifically, image features are extracted from Vision Transformer (ViT) [DOS20] and text features are extracted from BERT [DEV18], and are clustered together. If the clusters of two arbitrarily selected image features are different and the similarity between images calculated by ViT is low, and if the clusters between the given texts are also different, then the texts are exchanged to generate an incorrect pair. By repeating this process until the number of data is the same as that of the incorrect pair, only incorrect pairs can be generated without bias in the number of data and with low relevance of content between image texts. If the proposed method can successfully increase the accuracy of image retrieval, it will contribute to improving the purchasing effectiveness of online shopping by eliminating the need to filter out unwanted products.

This paper is organized as follows: Section 2 describes related work and their problems. Section 3 describes the proposed method, and Section 4 describes the experimental results of the proposed method and a comparative method. Section 5 provides a conclusion of the proposed method.

2 RELATED WORK

Various methods for image search have been proposed over time, traditionally utilizing tags attached to images or surrounding text content. Recently, multimodal approaches have become feasible, with methods proposed for searching images from input text using the embedding representations of BERT as metadata attached to images [QI20][LI23], as well as using CLIP and ALIGN [JIA21] for image search [BAL22][HEN22]. In fact, numerous image search methods have been proposed for CLIP thanks to its ability to learn semantic relationships between natural language text and image content.

For example, e-CLIP [SHI22] was designed for practical use in online shopping, proposing an image search

framework that utilizes CLIP for learning. It is aimed for use in downstream tasks such as category classification, attribute extraction, product matching, product clustering, and adult product recognition, allowing for the reduction of redundant information learning through the deletion of duplicate images using ResNet-34 [TAN19] and hash values, thereby enabling more efficient learning processes. By collecting similar images based on categories for contrastive learning, it has shown high performance in tasks related to the images used for learning, although it faces issues with decreased accuracy in zero-shot tasks.

The EI-CLIP [MA22] method improves the discriminative performance of images in texts containing proper nouns (e.g., Burberry, GUCCI) within the CLIP framework. It vectorizes proper nouns using an entity encoder and associates these vectors with textual embedding representation through an Entity-Aware module while contrastively learning with images. While it shows high discriminative performance for images in texts containing proper nouns, it faces issues with decreased accuracy in texts with high ambiguity or without metadata.

OpenFashionCLIP [CAR23], another method, seeks to enhance image search performance in the fashion domain not by modifying the learning method of CLIP, but by automatically performing prompt engineering on fashion-related texts to improve the quality of input queries. It prepares multiple template prompts and combines randomly chosen template prompts with input texts before feeding them to the text encoder for contrastive learning with images. Although it shows a higher discriminative performance than the baseline CLIP, selecting prompts randomly can result in prompts with low consistency with images, potentially leading to decreased discriminative performance depending on the dataset used.

There is also the RA-CLIP [XIE23], which, like the proposed method, improves the dataset based on the similarity between images. To enrich the information in the image data used to train the CLIP model, this method uses a module called RAM to extend the feature set of images that are similar to the input images. The training using these expanded image features significantly improves the zero-shot accuracy in image retrieval. Similar images play the role of a cheat sheet in CLIP, and the process of the proposed method can be regarded as an open-book test that does not require memorization of all visual information in the training data. Therefore, many image-text relationships can be learned with limited training. However, the real-time collection of similar images is computationally expensive, and retrieval performance degrades when the dataset is not domain- or task-specific.

Target learning by collecting similar images or by improving the quality of images and texts can improve retrieval accuracy to a certain extent. However, when images in the data are similar to each other, they are learned as low similarity of incorrect pairs with high relevance between image-text, and thus do not solve the essential problem. This method proposes a new method for generating good incorrect pairs.

3 PROPOSED METHOD

This section provides an overview of the proposed method and then describes the detailed method procedure.

3.1 Overview of proposed method

The proposed method generates effective incorrect pairs to improve the performance of image retrieval using CLIP. First, a new database (DB) is created from an original DB of image-text pairs (Figure3). The new DB contains correct pairs with strong semantic relevance and incorrect pairs with weak semantic relevance between images and texts, which are used for fine-tuning based on contrastive learning in CLIP. In vanilla CLIP, for K image-text pairs, the corresponding pair in a $K \times K$ matrix is treated as the correct pair, and the rest are learned as incorrect pairs. In contrast, the proposed method generates K incorrect pairs for each K correct pairs and performs contrast learning using these incorrect pairs. The method of generating incorrect pairs is described in Section 3.2, and the learning method using incorrect pairs is described in Section 3.3.

3.2 How to generate incorrect pairs

Unlike vanilla CLIP, the proposed method uses only incorrect pairs with low relevance between image-text pairs for training. First, features are extracted from the images and texts in the original DB shown in the Figure4 (left), using ViT and BERT, respectively, and clustered. In this case, we use K-means clustering [MAC67], which is reportedly effective for sentence modeling and topic modeling in previous studies [ASK21]. Next, one cluster that differs from the cluster of a certain image in the original DB is randomly selected, and the image features that are least similar to the image features in the original DB by Cosine similarity are output from the selected cluster. If the clusters between text features attached to the image are also different, the texts are exchanged. If the clusters between the texts are the same, the second and subsequent dissimilar image features are used to compare the text clusters. In this example, after the exchange, the image representing a jacket is assigned the text “shirt" and the image representing a shirt is assigned the text “jacket", as shown on the right in the

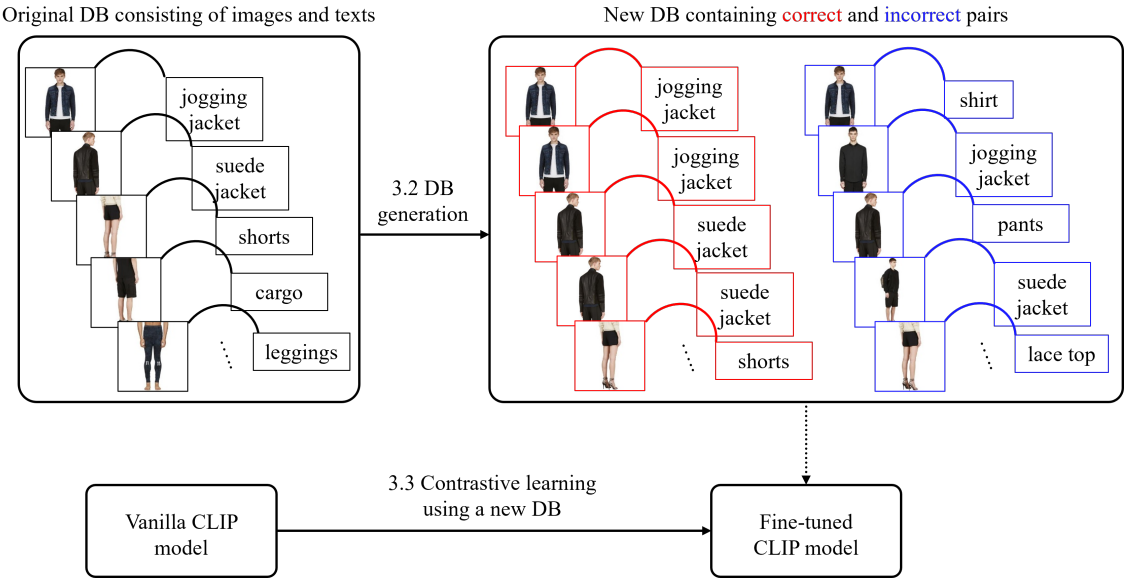

Figure 3: Flow of proposed method.

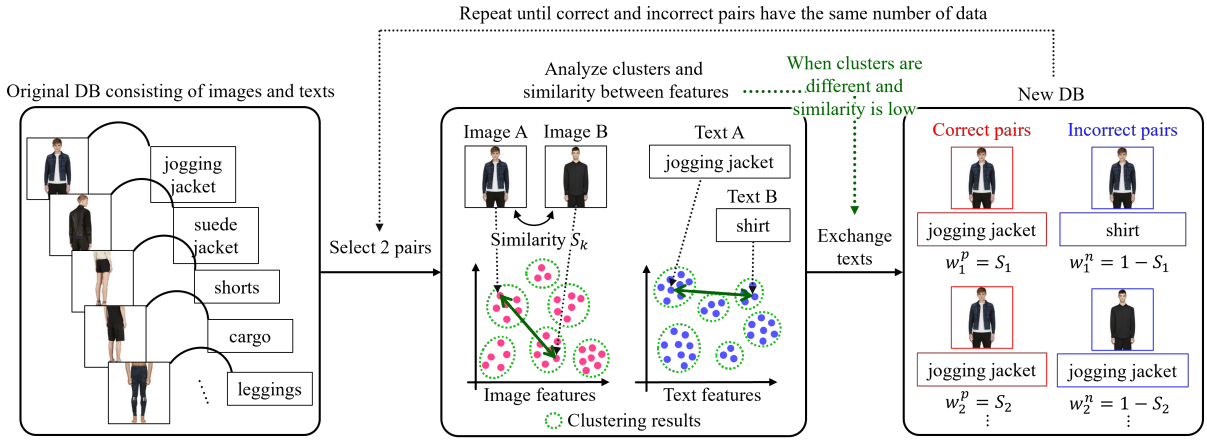

Figure 4: Flow of incorrect pair generation.

Figure. This is repeated until the number of correct pairs of data is reached.

Based on the assumption that "when there are multiple similar images, the texts assigned to the images are also similar to each other," we assign the weights w_k^p and w_k^n to each pair according to the similarity S_k between the two images for each pair (Eq.1, Eq.2). If the images are similar, the learning of correct pairs is enhanced by w_k^p during contrast learning, and if they are dissimilar, the learning of incorrect pairs is enhanced by w_k^n. In this way, the proposed method can separate only irrelevant images and text in the feature space without separating relevant images and text by learning with incorrect pairs generated by the proposed method.

$$w_k^p = S_k \tag{1}$$

$$w_k^n = 1 - S_k \tag{2}$$

3.3 Learning with incorrect pairs

The proposed method performs contrastive training utilizing correct and incorrect pairs with the same number of data. First, image and text features are extracted through encoders. Next, the similarity of the correct pair is trained to increase, while the similarity of the incorrect pair is trained to decrease (Figure 5). Utilizing the similarity between the correct and incorrect pairs of images as a basis, the training weights of the correct pair are balanced with the training weights of the incorrect pair. Specifically, when the similarity between the images of a pair is high, the training weight of the correct pair is increased; when the similarity is low, the training weight of the incorrect pair is increased. In this way, the model is expected to more effectively identify

relevant images based on the input text and reduce the output of irrelevant images. The loss is calculated from two terms, as in Eq.3. The first term L_p is the cross-entropy error between the cosine similarity s^p of the correct pair and the label y_k representing the category of the kth image, which is calculated by Eq.4.

The second term L_n is the cross-entropy error between the cosine similarity $1-s^n$ of the incorrect pair and the label y_k representing the category of the kth image, which is calculated by Eq.5. The similarity between the image and text features is subtracted from 1, so the less similar the features are, the smaller the loss. The lower the similarity between the correct pair of images and the incorrect pair of images, the smaller the value of w_k^p and the larger the value of w_k^n. α and β in Eq.3 are hyperparameters, which are weights that balance the terms. Thus, it can be seen that Learning with weights w_k^p, w_k^n can improve the discriminative performance of images because correct and incorrect pairs can be more distinctly separated in the feature space.

$$L = \alpha L_p + \beta L_n \tag{3}$$

$$L_p = -\sum_{k=1}^{K} y_k \cdot \log(w_k^p s_k^p) \tag{4}$$

$$L_n = -\sum_{k=1}^{K} y_k \cdot \log(w_k^n (1-s_k^n)) \tag{5}$$

4 EXPERIMENTS AND DISCUSSION

In this section, the purpose and conditions of the experiment are described, followed by the results and discussion of the experiment.

4.1 Experimental conditions

To evaluate the image retrieval performance of the proposed method, we conducted experiments on the task of searching for images from text. The ability to search and output the matching item (GT) from a group of images when a text query is provided is assessed. Following the experimental method of a prior study [GAO20], we utilized 100 randomly selected images from the dataset, conducting image searches using the text as a pair for the image. The experimental results are the average of 100 trials. The evaluation metrics are Rank@1, Rank@5, Rank@10, and Mean Average Precision (mAP). Rank@K indicates the percentage of correct images among the top K search results, and mAP is the average of how often the correct item appears at the top of the search results (AP) for each query. The datasets used were Fashion-gen [ROS18] and Amazon Review Data 2018 [NI19] (Figure 6). The Fashion-gen dataset contains 67,666 image-text pairs, of which 32,213 were used for the experiments. This dataset is characterized by its long and specific texts. The Amazon Review dataset contains 431,492 image-text pairs, with 30% of these being used in the experiments. This dataset provides a more challenging image search task due to its brief and more ambiguous texts compared to Fashion-gen. The number of clusters used for k-means clustering was set to 10, since the dataset is broadly classified into 10 categories.

As comparative methods, we used: (1) the CLIP model released by OpenAI [RAD21], (2) the CLIP model fine-tuned only with correct pairs, (3) the CLIP model fine-tuned with both correct and incorrect pairs (incorrect pairs were randomly selected from the DB without weighting w_k^p, w_k^n during training), (4) the CLIP model fine-tuned with both correct and incorrect pairs (incorrect pairs were generated from those with a similarity of 0.8 or less, without weighting w_k^p, w_k^n during training), (5) the CLIP model fine-tuned with both correct and incorrect pairs (incorrect pairs were generated from those with a similarity of 0.2 or less, without weighting w_k^p, w_k^n during training), (6) the CLIP model fine-tuned with both correct and incorrect pairs (incorrect pairs were generated from those with a similarity of 0.2 or less, with weighting w_k^p, w_k^n during training), (7) EI-CLIP (results cited from the paper [MA22]), and (8) OpenFashionCLIP [CAR23].

4.2 Quantitative experimental results

Table 1 lists the experimental results, with the vertical axis representing the evaluation metrics and the horizontal axis representing the methods, with the highest values for each metric highlighted in red. First, as a result common to both datasets, (1) had low generality for the data, resulting in low scores for all metrics, and (2) showed improved mAP scores compared to (1), but Rank@1 was low. For (3) and (4), there were cases where the images in the correct and incorrect pairs were similar, and there was no improvement in accuracy, but for (5) and (6), where the images in the correct and incorrect pairs were not similar, an improvement in mAP was observed. The proposed method generally showed high scores for Rank@1 and mAP, indicating that the clustering of image features and text features based on incorrect pairs is effective for learning. Additionally, adjusting the learning weights based on the similarity of images in correct and incorrect pairs, enhancing the similarity of correct pairs, and reducing the similarity of incorrect pairs all contributed to a clearer separation in the feature space and improved the image identification performance. As for (7), it was suggested in the original paper [MA22] that accuracy may decrease in cases without metadata or with high ambiguity of text, and it

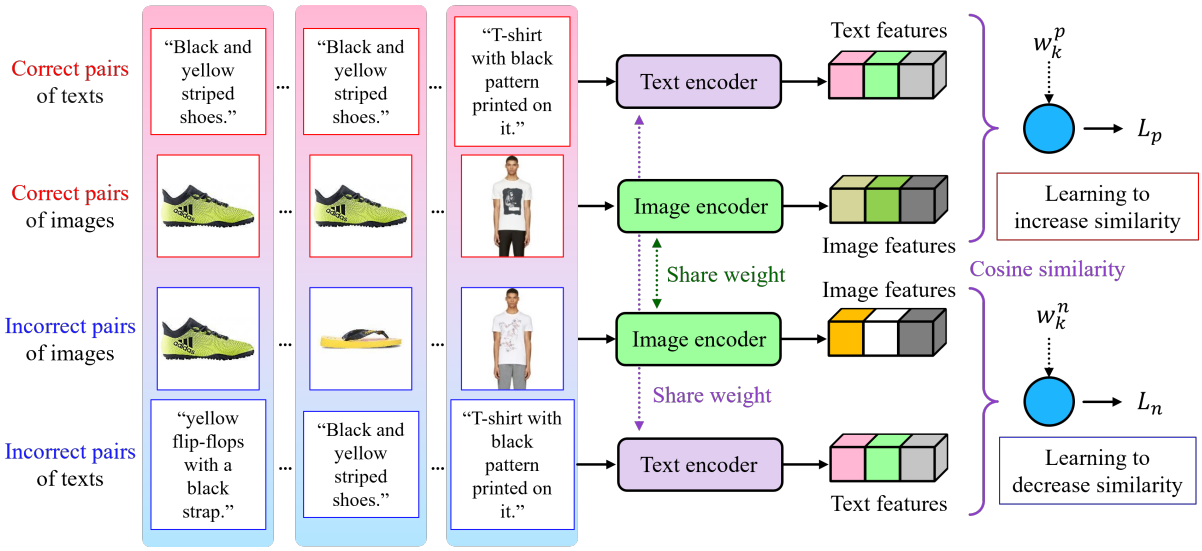

Figure 5: Learning flow.

Fashion-gen dataset

	(1)	(2)	(3)	(4)	(5)	(6)	Ours	(7)	(8)
Rank@1	44.0	60.0	47.0	65.0	64.0	65.0	71.0	40.0	54.0
Rank@5	81.0	93.0	87.0	92.0	95.0	96.0	95.0	71.0	83.0
Rank@10	94.0	100	95.0	99.0	99.0	100	97.0	84.0	88.0
mAP	0.60	0.73	0.64	0.77	0.78	0.78	0.81		0.65

Amazon Review Data 2018 dataset

	(1)	(2)	(3)	(4)	(5)	(6)	Ours	(7)	(8)
Rank@1	59.0	59.0	62.0	64.0	67.0	71.0	72.0	23.7	56.0
Rank@5	80.0	81.0	83.0	83.0	82.0	84.0	84.0	49.4	78.0
Rank@10	85.0	89.0	89.0	89.0	87.0	89.0	90.0	61.6	84.0
mAP	0.67	0.68	0.70	0.72	0.73	0.77	0.78		0.66

Table 1: Results using Fashion-gen dataset (above) and results using Amazon Review Data 2018 dataset (below).

(a) Example image and text from Amazon Review Data 2018 dataset

(b) Example image and text from Fashion-Gen dataset

Figure 6: Example images and texts from datasets.

was observed that Rank@1 was low for the Amazon review data, which is composed of highly ambiguous text. (8) was considered to have not high identification accuracy due to the potential loss of the model's generalization capability for the data, which could be impaired by prompt engineering dependent on the method.

4.3 Qualitative experimental results

To qualitatively evaluate the image identification performance of the proposed method, the top 10 search images for each method were output in response to text queries. The dataset used was Digikala Products Color Classification [MAS21], which includes various similarly colored and shaped product images for online shopping, created for product identification. It consists solely of images, with no correct texts provided; therefore, the determination of whether the output images for the text queries were correct was based on subjective judgment and the judgment of ChatGPT-4. The comparative methods were all the same ones as used in Section 4.2 except for (7).

Figure 7 shows the results, with the vertical axis representing each method and the horizontal axis representing the output image. The red box in the image shows the image judged to be the correct image by ChatGPT-4 and subjectivity, and the black box shows the image not judged to be correct by ChatGPT-4. In this experiment, the text "Black and white striped shoes." was input and

Figure 7: Output results (top 10 images) when the text “Black and white striped sneaker." was input to each method.

the image was output. The experimental results suggest that the image identification performance was low for (1) and (2) because the training may have used incorrect pairs in which the image and text were highly related. In the case of (3) and (4), where the correct pair images may be similar or identical to the incorrect pair images, the training was not effective because the similarity between the image and text was trained to be low, even though the images and text were related. In the case of (5) and (6), the images with low similarity between the incorrect and correct pairs were selected, so we conclude that the intended image for the input text was output. In particular, since the proposed method clusters images and text independently, it is easy to understand the structure of the entire dataset and to select pairs that are not particularly similar, and as a result, we believe that the intended images can be output.

5 CONCLUSION

In this work, we proposed an incorrect pair generation method based on image clustering and demonstrated through experiments that it achieves a higher search accuracy compared to other incorrect pair generation methods and image search techniques using CLIP. By generating suitable incorrect pairs from the clustering results of image and text features and learning from them, the image identification capability was enhanced. Experimental results from the Amazon Review Data 2018 dataset, a commonly used dataset in this field, showed a 27.0% improvement in Rank@1 score compared to vanilla CLIP, and a 11.0% improvement compared to a random reduction of incorrect pairs. An image search system utilizing the proposed method would be able to save users the trouble of filtering out undesired products, thus improving usability and potentially enhancing the purchasing effect in online shopping. The proposed method only generates incorrect pairs of images and text that are too poorly related, which may lead to over-learning. In the future, we plan to increase the accuracy of image retrieval by learning the relationships among detailed features of objects from the relationships among image texts at a rough category level by increasing the learning hierarchy of the CLIP model. We will also expand the comparison method, encoders used, and datasets to further demonstrate the effectiveness of the proposed method.

6 REFERENCES

[SIG08] B. Sigurbjornsson and R. Van Zwol, “Flickr tag recommendation based on collective knowledge", In Proceedings of the 17th International World Wide Web Conference, pp. 327–336, Apr. 2008.

[KIL16] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, “Re-evaluating automatic metrics for image captioning", arXiv preprint arXiv:1612.07600., 2016.

[GIA15] S. Giannoulakis and N. Tsapatsoulis, “Instagram hashtags as image annotation metadata", In Artificial Intelligence Applications and Innovations: 11th IFIP WG 12.5 International Conference, AIAI 2015, Proceedings 11, pp.206–220, Springer International Publishing, Sep. 2015.

[QI20] D. Qi, L. Su, J. Song, E. Cui, T. Bharti and A. Sacheti, “Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data", arXiv preprint arXiv:2001.07966, 2020.

[LI23] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models", In International conference on machine learning, pp.19730–19742, PMLR, 2023.

[JIA21] C. Jia, Y. Yang, Y. Xia, Y. T. Chen, Z. Parekh, H. Pham, et al., “Scaling up visual and vision-language representation learning with noisy text supervision", In International conference on machine learning, pp.4904–4916, PMLR, 2021.

[RAD21] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, ... and I. Sutskever, “Learning transferable visual models from natural language supervision", In International Conference on Machine Learning, pp.8748–8763, PMLR, July 2021.

[BAL22] A. Baldrati, M. Bertini, T. Uricchio, and A. Del Bimbo, “Effective conditioned and composed image retrieval combining clip-based features", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.21466–21474, 2022.

[HEN22] M. Hendriksen, M. Bleeker, S. Vakulenko, N. van Noord, E. Kuiper and M. de Rijke, “Extending CLIP for Category-to-image Retrieval in E-commerce", In European Conference on Information Retrieval, pp.289–303, April 2022.

[AGA21] S. Agarwal, G. Krueger, J. Clark, A. Radford, J. W. Kim, and M. Brundage, “Evaluating CLIP: Towards characterization of broader capabilities and downstream implications", arXiv preprint arXiv:2108.02818., 2021.

[SHA23] J. J. Shao, J. X. Shi, X. W. Yang, L. Z. Guo, and Y. F. Li, “Investigating the limitation of clip models: The worst-performing categories", arXiv preprint arXiv:2310.03324., 2023.

[DOS20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16×16 words: Transformers for image recognition at scale", arXiv preprint arXiv:2010.11929, 2020.

[DEV18] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding", arXiv preprint arXiv:1810.04805, 2018.

[SHI22] W. Shin, J. Park, T. Woo, Y. Cho, K. Oh, and H. Song, “e-clip: Large-scale vision-language representation learning in e-commerce", In Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp.3484–3494, Oct. 2022.

[TAN19] Y. Tang, F. Borisyuk, S. Malreddy, Y. Li, Y. Liu, and S. Kirshner, “MSURU: Large scale e-commerce image classification with weakly supervised search data", In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp.2518–2526, July 2019.

[MA22] H. Ma, H. Zhao, Z. Lin, a. Kale, Z. Wang, T. Yu, et al., “Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.18051–18061, 2022.

[CAR23] G. Cartella, A. Baldrati, D. Morelli, M. Cornia., M. Bertini, and R. Cucchiara, “OpenFashionCLIP: Vision-and-Language Contrastive Learning with Open-Source Fashion Data", In International Conference on Image Analysis and Processing, pp.245–256, Sep. 2023.

[XIE23] C. W. Xie, S. Sun, X. Xiong., Y. Zheng, D. Zhao, and J. Zhou, “Ra-clip: Retrieval augmented contrastive language-image pre-training", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.19265–19274, 2023.

[MAC67] J. MacQueen, “Some methods for classification and analysis of multivariate observations", In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol.1, No.14, pp.281-297, 1967.

[ASK21] P. M. A. Kumar, T. S. M. Rao, L. A. Raj, and E. Pugazhendi, “An efficient text-based image retrieval using natural language processing (NLP) techniques", In Intelligent System Design: Proceedings of Intelligent System Design: INDIA 2019, pp.505–519, 2021.

[GAO20] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, ... and H. Wang, “Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval", In Proceedings of the 43rd International ACM SIGIR Conference on Research and Devel-

opment in Information Retrieval, pp.2251–2260, July 2020.

[ROS18] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, and C. Pal, “Fashion-gen: The generative fashion dataset and challenge", arXiv preprint arXiv:1806.08317, 2018.

[NI19] J. Ni, J. Li, and J. McAuley, “Justifying recommendations using distantly-labeled reviews and fine-grained aspects", In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pp.188–197, 2019.

[MAS21] masouduut94, 2021, “Digikala Products Color Classification | Kaggle", https://www.kaggle.com/datasets/masouduut94/digikala-color-classification/data