Product Matching with Contrastive Language-Image Learning

Shuyu Guo
MIT
shuyu555@mit.edu

Qinwen Wang
MIT
qinwenw@mit.edu

Abstract

Product matching is considered to be a crucial task in e-commerce, enabling retailers to identify and compare similar products across multiple platforms. This project aims to apply Convolutional Neural Network (CNN), Supervised Contrastive Learning, and Contrastive Language-Image Pretraining (CLIP) to perform product classification, matching, and recommendation tasks. We also propose a new method - Supervised Contrastive Language-Image Pretraining (supervised CLIP) that incorporates image and text embeddings and utilizes supervised contrastive loss function. Supervised CLIP is proven to be effective for product matching tasks given its highest top-k accuracy among the methods. The proposed approach has the potential to bridge the gap between customer needs and product supply, enabling a more personalized and efficient shopping experience.

1. Introduction

In today's fast-paced world, online shopping has become the preferred method for consumers seeking convenience and variety. The total e-commerce sales for 2022 were estimated at $1,034.1 billion, an increase of 7.7 percent (±0.4%) from 2021 [1]. As e-commerce platforms continue to grow, the sheer volume of products available can be overwhelming, making it difficult for shoppers to find the exact item they're looking for. It is therefore crucial for e-commerce platforms to help customers find the exact or similar items of interest from the inventory base. Product matching is challenging considering the large amount of product inventory and how different merchants present their products. This project aims to bridge the gap between customer needs and product discovery.

Recent works in product matching include using CNN for image analysis and using contrastive representation learning to find similar images. In this project, we first used pre-trained ResNet-50 to classify product images. Then we implemented the supervised contrastive representation learning, which generates feature embeddings for the product images and then uses a supervised contrastive loss function to identify similar products. We also explored CLIP, which generates embeddings for both images and text descriptions, and is designed to recommend relevant products given text input.

Inspired by supervised contrastive learning and CLIP, we propose a new multi-modal approach, supervised contrastive language-image pre-training (Supervised CLIP), which combines both image and text data and utilizes the supervised contrastive loss function to find similar products.

We experimented on a dataset provided by Shopee[1], a leading e-commerce platform in Southeast Asia and Taiwan. The dataset contains 34,250 product images with text descriptions and 11,014 unique label groups. Our proposed supervised CLIP approach outperforms the other methods with a top-1 accuracy of 83.01%, which is 5.45% higher than the supervised contrastive learning. Our main contributions are summarized below:

- We propose a new multi-modal approach that combines both image and text data and utilizes the supervised contrastive loss function to find similar products.
- We show that our proposed supervised CLIP approach provides a top-k accuracy boost from existing approaches.
- We perform experimentations on real-world e-commerce data, which provides a streamlined product matching process from data pre-processing, feature encoding, model training, and performance evaluation, which is simple to implement and deploy in a production scale.

2. Related Work

Our work draws on existing literature in supervised contrastive learning and CLIP.

[1] Data Source: https://www.kaggle.com/competitions/shopee-product-matching

2.1. Supervised Contrastive Learning

In recent years, Supervised Contrastive Learning [2] has gained considerable attention for its success in learning effective feature representations from large datasets, particularly in the field of computer vision. It has been applied to product matching in the past few years since its object, seeking to learn representations that remain invariant across diverse transformations of a single instance while demonstrating discriminative abilities against distinct instances, aligns with the product matching task. Peeters et al. [3] apply supervised contrastive learning to product matching in e-commerce using product offers from different online shops. The study concludes that contrastive pre-training holds significant potential for product-matching use cases where explicit supervision is available.

2.2. CLIP

For most e-commerce platforms, customer input desired product descriptions and expect a list of recommendations. Therefore, developed by OpenAI, contrastive language-image pretraining (CLIP) [4], a combination of language model and computer vision might also be a good candidate for product matching. The key idea behind CLIP is to train a neural network to predict the correspondence between images and texts, allowing it to understand both modalities. The model is trained using a contrastive learning objective, which encourages the model to produce similar representations for image-text pairs that are semantically related while producing dissimilar representations for unrelated pairs. As detailed in the research by Hendriksen et al. [5], the application of CLIP in the context of e-commerce has been successful, particularly in the task of category-to-image retrieval. Their model, CLIP-ITA, leverages the rich, multimodal data available in e-commerce, including visual, textual, and attribute modalities, to create comprehensive product representations.

3. Methodology

3.1. Dataset

The dataset provided by Shopee contains 34,250 product images with text descriptions and 11,014 unique label groups. As shown in Figure 1, each label group contains on average three similar products. The original images have a size of (1024, 1024, 3), and we decide to use a 40 batch size, and 512 as our input size. The product titles include both English and non-English text descriptions, and the lengths are varying by product. The text descriptions often include the specific product brand and product name, i.e. 'Paper Bag Victoria Secret'.

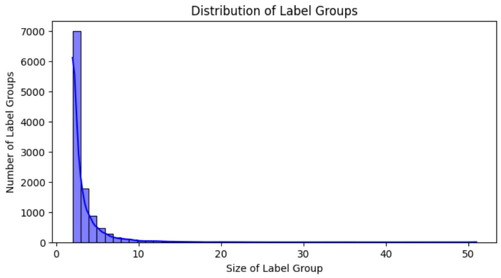


Figure 1. Frequency distribution of label groups in Shopee dataset: The data contains 11,014 unique label groups. Most label groups include less than 10 products in each group. Only 0.53% label groups contain more than 20 images.

3.2. Pre-processing

For product images, we performed a series of image transformations in the pre-processing step, including a horizontal flip with $p = 0.5$, a random brightness contrast with brightness limit $(-0.1, 0.1)$, contrast limit $(-0.1, 0.1)$, $p = 0.5$, and a normalization.

For product titles, we employed several natural language processing (NLP) techniques for the pre-processing. We removed punctuations and common English stopwords provided by `nltk`. We also removed stopwords specific to product descriptions, including numbers and measurement units. We did lemmatization using WordNetLemmatizer. Finally, we performed tokenization using DistilBertTokenizer.

3.3. Proposed Methods

Convolutional Neural Network. In this project, we explored various Convolutional Neural Network (CNN) architectures for product image classification, including the well-known Residual Network (ResNet). We first tried fine-tuning the pre-trained ResNet-50 [6] model using a small subset of the product images and classifying images into 11,014 groups based on their actual label groups. However, the accuracy is significantly poor since there are too many classes and each class only has three images on average.

Supervised Contrastive Learning. Apart from traditional CNN, we also implemented Supervised Contrastive Learning [2], as its objectives align well with product matching tasks. We chose to use supervised over self-supervised because we aim to maintain the proximity of similar instances (e.g., handbags and totes) in the learned feature space by using the provided labels. We believe customers may still be interested in similar but not identical products. Figure 2 demonstrates the difference between supervised vs. self-supervised contrastive learning methods.

We learned the image embeddings using the tf-

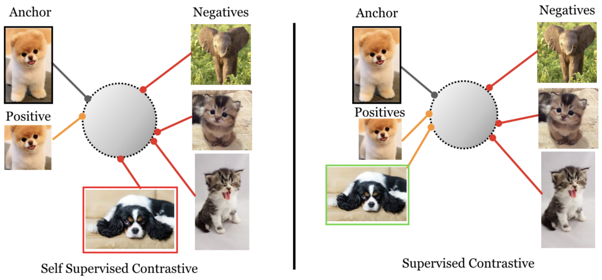


Figure 2. Comparison between self-supervised contrastive and supervised contrastive losses [2]: As illustrated by the black and white puppy image, we observe that when incorporating class label information, the supervised loss is more accurate in terms of bringing image features for the same class closer compared to the self-supervised loss.

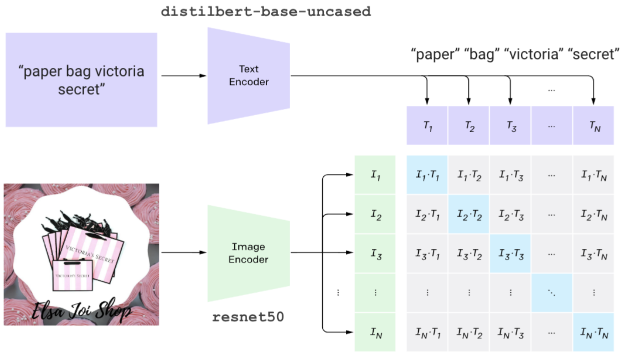
efficientnet-b4-ns model [7], a pre-trained EfficientNet model, which is a state-of-the-art convolutional network for image classification tasks, designed to provide excellent accuracy while also being computationally efficient.

We use supervised contrastive loss as the loss function, as proposed by Khosla et al. [2]. As shown in the equation below, the supervised contrastive loss contrasts the set of all samples from the same class as positives against the negatives from the remainder of the batch.

$$L^{con} = \sum_{i\in I} \frac{-1}{|P(i)|} \sum_{p\in P(i)} \log \frac{exp(z_i \cdot z_p/\tau)}{\sum_{a\in A(i)} exp(z_i \cdot z_p/\tau)} \tag{1}$$

Here, the index i is called the anchor, and the index p represents the positive that is in the same label group as i. $A_i = \{I \setminus i\}$ is the set of indices in the images distinct from i. $P(i) = \{p \in A(i) : \tilde{y_p} = \tilde{y_i}\}$ is the set of indices of all positives in the multiview batch (all transformed versions of the images) distinct from i, and $|P(i)|$ is its cardinality. z_i represents the feature vector of i in the projection space. τ is a scaler temperature parameter. In this project, we used $\tau = 0.1$.

CLIP. The CLIP model is pre-trained on a large dataset of paired image and text data, which enables it to learn rich representations of both modalities. With the pre-training contrastive learning approach, CLIP involves training the model to distinguish between positive pairs of images and text descriptions and negative pairs, where the text description does not match the corresponding image. As shown in Figure 3, in this project, we chose the DistilBERT-base-uncased model to generate text embeddings and the ResNet-50 model for image embeddings. We believe that the model's ability to seamlessly bridge the gap between visual and linguistic representations is a powerful asset, particularly in tasks requiring a nuanced understanding of both modalities, such as product matching in e-commerce.



	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

Figure 3. CLIP framework: We used distilBERT-base-uncased as the text encoder and ResNet-50 as the image encoder to learn the representations of product titles and images. We then calculated the cosine similarity and used the contrastive loss function to train the embeddings.

The fundamental mechanism driving CLIP is based on a contrastive learning paradigm. More formally, given a set of paired image and text data (x_i, y_i), the model is trained to maximize the similarity between the image and text pair (x_i, y_i) while minimizing the similarity with negative samples. This can be captured by the following contrastive loss function:

$$L^{CLIP} = -\sum_{i} \log \frac{\exp(f(x_i)^T g(y_i)/\tau)}{\sum_{j\neq i} \exp(f(x_i)^T g(y_j)/\tau)}, \tag{2}$$

where $f(x_i)$ and $g(y_i)$ are the image and text representations, respectively, and τ is a temperature parameter.

Supervised CLIP. Finally, we leveraged both text (product title) and image information from the dataset and developed a multi-model approach - supervised contrastive language-image learning. We believe that the proposed method utilizes the multimodal nature of product matching tasks and will outperform the other methods.

As depicted in Figure 4, we concatenated the text embeddings from the distilBERT-base-uncased encoder and image embeddings from the ResNet-50 encoder together to form a combined representation for each product, so that the model is able to learn how to match products using both text and image information. This combined representation is then used to compute the similarity between different products.

The loss function for supervised CLIP is similar to that (Equation 1) of Supervised Contrastive Learning with $z_i =$ Concat$(f(x_i), g(y_i))$ instead, where z_i represents the final concatenated embedding for the ith product, $f(x_i)$ is the image embedding, and $g(y_i)$ is the text embedding.

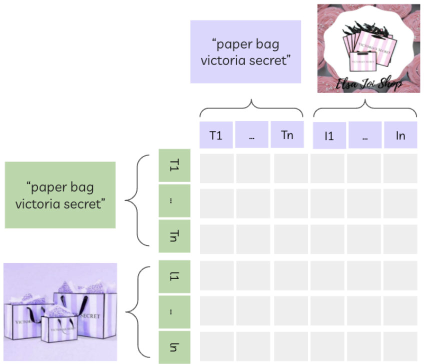


Figure 4. Supervised contrastive language-image learning framework: We used distilbert-base-uncased as the text encoder and tf-efficientnet-b4-ns as the image encoder to learn the representations of product titles and images. We then calculated the cosine similarity and used the supervised contrastive loss function to train the embeddings.

3.4. Performance Evaluation

To evaluate and compare the performance of different methods, we defined a unified evaluation metric - top k matching accuracy. We first extracted the feature embeddings (image embedding for supervised contrastive learning, image and text embeddings for CLIP and supervised CLIP) for the validation set from the best-trained model. Note that we used a unified embedding size of 512 for all methods. We then calculated the cosine similarities between each pair of products in the validation set and selected top k (k=1,2,5) similar matches for each product. The accuracy is calculated by comparing the top k matches against the true matches provided by the label groups in the dataset. Since there might be more than one item that is matched with a product, we consider the matching to be accurate if there is one true match in the k items.

This metric provides a more forgiving measure of the model’s predictive capabilities, as it allows for the correct match to be within the top k predictions rather than strictly at the top. This is especially relevant for our product matching scenario, where providing a shortlist of potential matches could be equally valuable as identifying the exact match.

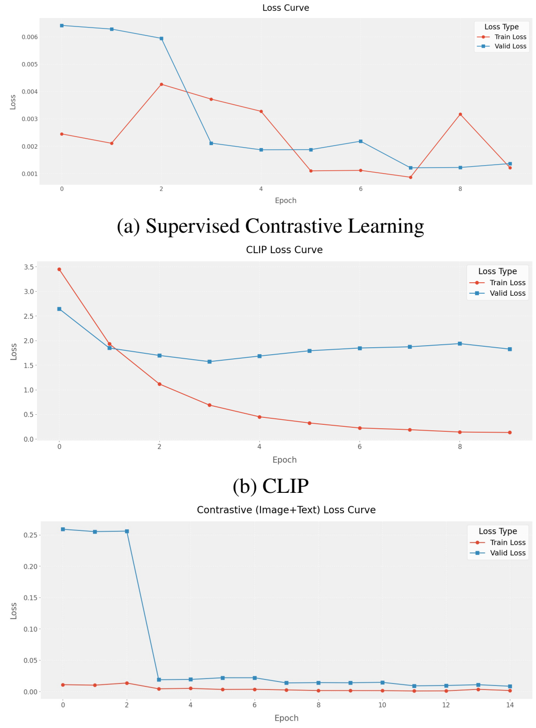
4. Result

4.1. CNN Result

Given the large number of unique groups in the dataset, traditional Convolutional Neural Networks (CNNs) is not the optimal candidate to categorize two to three images into a single group. In our preliminary experiment, we observed that the best validation accuracy achieved after 10 epochs on a subset of the entire dataset was only 1.01 %. Although we anticipate that the performance might improve by increasing the number of training set and fine-tuning the hyperparameters, we believe that this approach is not suitable for product matching tasks. Therefore, we decided to focus on contrastive representation learning and CLIP for the task.

4.2. Loss Curve

We evaluate the training losses for the three methods, as shown in Figure 5.



(a) Supervised Contrastive Learning

(b) CLIP

(c) Supervised CLIP

Figure 5. Loss curve comparison: Supervised contrastive learning reaches a valid loss of 0.0014 at epoch=5. CLIP reaches a valid loss of 1.8 at epoch=10. Supervised CLIP reaches a valid loss of 0.0084 at epoch=15. We observe an overfitting effect on CLIP. Note that the losses between models are not directly comparable because they are evaluated with different metrics (the losses in Equation 1 and in Equation 2).

4.3. Top-k Accuracy

We show the top-k (k=1, 2, 5) matching accuracy for the three methods in Table 1. The models using the supervised contrastive loss function (Equation 1) have a better top-k accuracy compared to the one using the cross-entropy loss (Equation 2). This outcome is in line with anticipated results, as supervised contrastive losses tend to leverage labeled data more effectively, thereby facilitating improved predictive performance. Furthermore, when both text and image information are incorporated into the learning pro-

	Supervised Contrastive Learning	CLIP	Supervised CLIP
k=1	77.56%	34.40%	**83.01**%
k=2	82.25%	41.77%	**88.05**%
k=5	87.29%	46.31%	**92.61**%

Table 1. Top-k matching accuracy: Supervised CLIP outperforms supervised contrastive learning by 5.45% and CLIP by 48.61% in terms of top-1 accuracy.

cess, the supervised CLIP model outperforms all other models. This finding reinforces the premise that integrating multiple data modalities can lead to enhanced model performance. It suggests that a more comprehensive understanding of the product can be achieved when both its visual and textual descriptors are taken into account, leading to more accurate product matching outcomes.

CLIP is not effective in product matching tasks according to the accuracy results because its primary design is to establish correspondence between text descriptions and their related images, rather than focusing on matching similar products. The model is able to predict general categories but hard to identify more granular details like brands, styles, etc. While it may not be as effective for exact product matching, it showcases the strong potential for applications in product recommendation systems (demo for product recommendation shown in Figure A.2). This is because consumers often display interest in exploring a variety of products that share similarities or 'lookalike' traits, instead of being exclusively interested in finding an exact product. Thus, while CLIP's top-k accuracy may be lower compared to supervised contrastive learning for exact product matching, it may still offer valuable capabilities for related tasks in the e-commerce domain.

5. Discussion & Conclusion

5.1. Conclusion

In conclusion, the supervised contrastive loss function excels in the task of product matching especially when training with both text and images as input, owing to its inherent design of learning representations that are similar for the same instances while being discriminative against different instances. This characteristic aligns well with product matching, where the goal is to identify similar products within a large dataset. On the other hand, CLIP's strength lies in its ability to understand both image and text modalities. We observe that while CLIP is not as effective for exact product matching, it displays considerable potential for product recommendation systems.

This project presents compelling evidence that machine learning techniques that leverage multi-modal learning (i.e., supervised CLIP) can enhance the task of product matching in e-commerce. The findings of this study provide valuable insights for future research and development in this field, with potential applications not only in product matching but also in product recommendation and customer personalization.

5.2. Future Works

We discuss our main limitations and future work as follows.

On training/validation split We observe that the train/valid split affects the performance of the learning as there are imbalances in label groups. Some label groups have 50 similar products whereas most label groups only have two to three products. Future work related to applying optimization to find an optimal train/valid split can be considered.

On model configurations We observe that the top-k accuracy is affected by model configurations, i.e., the choice of text/image encoders (ResNet or EfficientNet), learning rates, embedding size, etc. The configurations can be examined more closely and more experimentations can be conducted.

On multimodal product matching There are a lot of prevalent multimodal methods for product matching in e-commerce. Besides concatenation, which is what we did in the project, we can also explore the attention mechanism [8], Siamese Network [9], and VisualBERT [10] to fuse the image and text embeddings.

On ethics Machine learning models can inadvertently learn and perpetuate biases present in the training data. For instance, if a model is trained predominantly on some particular brands or products, it may not perform as well when matching other brands or products. This could lead to unfair outcomes for certain users or vendors. We can explore various data resampling techniques to overcome class imbalance issues and reduce bias.

References

[1] U.S. Census Bureau, "Quarterly retail e-commerce sales, 4th quarter 2022," Retrieved from https://www.census.gov/retail/mrts/www/data/pdf/ec_current.pdf, 2023. 1

[2] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," 2021. 2, 3

[3] R. Peeters and C. Bizer, "Supervised contrastive learning for product matching," apr 2022. [Online]. Available: https://doi.org/10.1145%2F3487553.3524254 2

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. 2

[5] 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," Cham, pp. 289–303, 2022. 2

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," https://arxiv.org/abs/1512.03385, 2015. 2

[7] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *CoRR*, vol. abs/1905.11946, 2019. [Online]. Available: http://arxiv.org/abs/1905.11946 3

[8] N. Das, A. Joshi, P. Yenigalla, and G. Agrwal, "Maps: Multimodal attention for product similarity," 2022. [Online]. Available: https://www.amazon.science/publications/maps-multimodal-attention-for-product-similarity 5

[9] K. Gupte, L. Pang, H. Vuyyuri, and S. Pasumarty, "Multimodal product matching and category mapping: Text+image based deep neural network," pp. 4500–4505, 2021. 5

[10] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang, "Visualbert: A simple and performant baseline for vision and language," 2019. [Online]. Available: http://arxiv.org/abs/1908.03557 5

Appendix

A. Figures

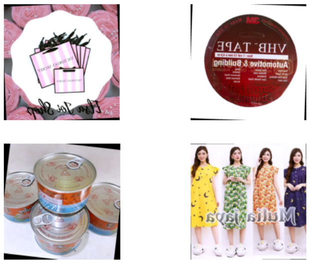


Figure A.1. Examples of transformed images: We performed a series of image transformations including horizontal flip, random brightness contrast, and normalization.

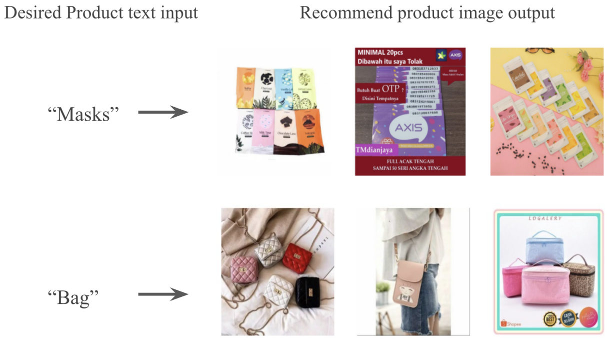


Figure A.2. CLIP product recommendations demo: When we input keywords like "Masks" or "Bag", CLIP retrieves relevant product images (shown on the right). Note that given the example of "Bag", CLIP can generate accurate images in terms of general categories, but is less efficient in retrieving specific brands or styles.