Say Yes To The Fetch:
Product Retrieval on a Structured Multimodal Catalog

*

João Bernardo Morais, Carlos Santiago, João Paulo Costeira

Instituto Superior Técnico

{joao.bernardo.morais, carlos.santiago, jpcosteira }@tecnico.ulisboa.pt

Abstract

Multimodal conversational agents allow the user to communicate through natural language and visual information. In e-commerce, this type of agents have the potential to lead to realistic and dynamic shopping experiences, where the costumer finds the desired products more efficiently with the help of an agent. A common approach in this scenario is to build a representation space where both the textual and visual information of a product are close. Then, it is possible to search and retrieve products with queries from any of the modalities. This work proposes to generate this joint representation space by also taking into account prior knowledge about the fashion domain, to ensure that the retrieved products comply with the target type of products. Combining label relaxation with a taxonomy-based regularization, the proposed approach diminishes the penalization of the contrastive loss by assigning a smaller loss to other acceptable matches. Our results show that the proposed approach significantly reduces gross errors, like retrieving pants when the costumer is looking for t-shirts, while simultaneously achieving good retrieval performances. Additionally, this approach allows multimodal queries, where specific attributes can be modified by manipulating a visual query with text.

Introduction

Nowadays, with e-commerce rising in completely different areas, the need to incorporate the in-store physical experience in online shopping rapidly emerged. One of the ways to accomplish this is through conversational agents capable of searching and retrieving the products desired by the user from the store's catalog. For an efficient interaction, especial attention has been given to multimodal conversations, where the user and the agent communicate through both textual and visual data. This type of interaction leads to realistic, fast, and dynamic experiences, with the potential to revolutionize e-commerce (Magalhaes et al. 2021).

This work aims to provide a conversational agent the ability to perform product retrieval based on a query given during the conversation The query can be given in a textual or visual format, to allow taking advantage of all the different types of data that fully characterize fashion items. A difficulty that immediately arises is the challenge of representing different objects that contain the same or similar information.

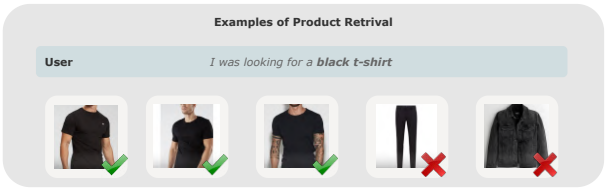

Figure 1: Right and wrong examples of product retrieval.

Within the fashion domain, products are characterized not only by their visual aspect and style, but also by its category, attributes, brand, etc. Capturing all this information in a single representation is crucial for the success of product retrieval (Li et al. 2021), in particular to prevent critical failures, as suggested in Figure 1.

Thereupon, arises the need to create a common space where both modality instances coexist in the form of the representations generated from the information extracted from the products. This work proposes a new label relaxation strategy to train a retrieval model based on a structured multimodal space. Instead of minimizing the contrastive loss between the corresponding textual and visual embeddings, as in (Radford et al. 2021), our approach relies on prior knowledge extracted from the products to impose additional constraints on the organization of the representation space.

Our proposed approach assumes that multiple valid matches may exist between the textual and visual instances. Namely, products that share a similar path in the fashion taxonomy are assumed to be close matches and are, therefore, less penalized in the loss function. This new approach originates a structured representation space, represented in Figure 2, that captures the nuances of product retrieval, while minimizing critical errors.

Related Work

Several works have been developed on product retrieval. In a general way, these works extract features from the different

*With help from the AAAI Publications Committee. This work was supported by LARSyS - FCT Project UIDB/50009/2020 and project iFetch, Ref. 45920, co-financed by ERDF, COMPETE 2020, NORTE 2020 and FCT under CMU Portugal.

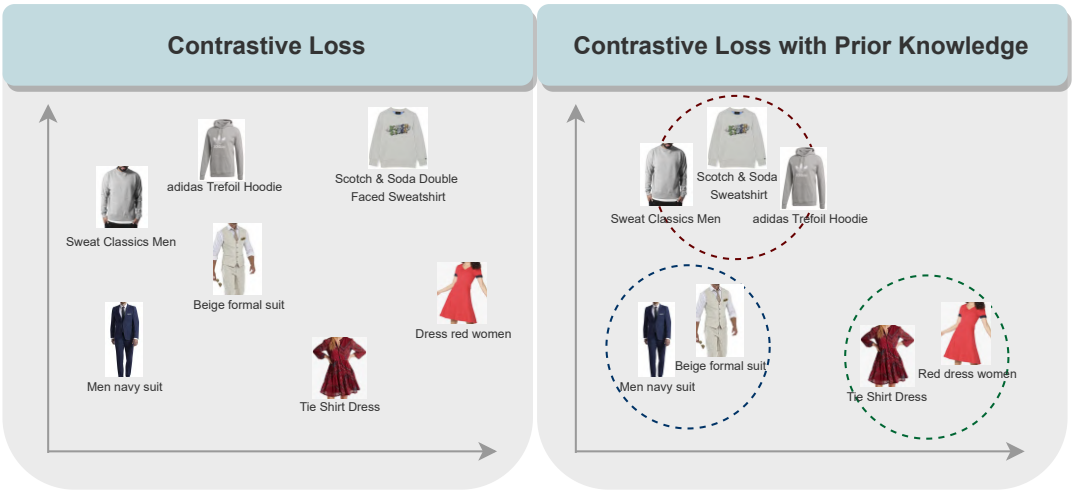

Figure 2: Spatial organization with a contrastive loss versus the proposed relaxed contrastive loss with prior knowledge.

objects, and learn to match queries with the corresponding products through metric learning. Van Gysel, de Rijke, and Kanoulas (2016) introduced a latent vector space model that learns representations of words, e-commerce products (that are associated with its respective textual description and at least one user review), and a projection between the two of them. This latent representation enables retrieval based on conceptual content instead of exact word matching. Rasiwasia et al. (2010) used Canonical Correlation Analysis to study the correlations between two modalities in multimedia documents in order to find a projection to connect them and enable cross-modal document retrieval, i.e, fetching the text that reflects a given image, or fetching the image that reflects a textual query. Gong et al. (2013) extended the previous approach in order to overcome some weaknesses of this method, such as the incapacity to integrate additional information that would lead to a supervised learning approach and a space structuring by introducing a third dimension. Different works have also tried to solve some computer vision tasks using NL, by connecting the representation of images and text. In one of the first approaches of trying to label images, Hironobu, Takahashi, and Oka (1999) proposed a method that creates relationships between images and words. The ambition is the ability to detect words (nouns and adjectives that would describe or summarize the image) in paired images. DeViSE (Frome et al. 2013) is another architecture that ends up bridging visual and textual representations. The goal is to, again, transfer textual semantics into a model trained for visual object recognition by connecting two modalities, but this time by introducing a convolutional neural network and with the ambition of achieving the ability to perform zero-shot transfer. Considering a dataset of both labeled image data and raw unannotated text, the authors concluded that semantic information demonstrated to be useful to make predictions about thousands of image labels not observed during training. Focusing on image captioning, Li et al. (2017) extended the task of predicting individual words present in the image's caption into a much more complex task: predicting phrase n-gram that can be seen as a caption with length n using the same dataset as before. In particular, the authors considered for this assignment images and respective user comments excluding labeled data. For this, visual n-grams models that can formulate random phrases relevant to the content of an image are built.

Baltrušaitis, Ahuja, and Morency (2017) proposed two categories of multimodal representation such that it is possible to distinguish different ways of connecting different objects. One of them, a *coordinated representation*, generates a coordinated space by creating similarity constraints that combine the individual modality's signals after being individually processed. For the generation of this coordinated representation, a visual and textual encoders are used to generate the visual and textual representations, respectively, that are later approximated with the cosine-similarity.

Recently, CLIP (Radford et al. 2021) has surpassed the previous works by considering a contrastive approach in which the goal is to match raw captions with images. The model is able to perform zero-shot transfer to downstream tasks by using Natural Language as a supervisor that expresses an enormous number of visual concepts and has shown to be competitive with fully-supervised baselines.

Nevertheless, due to its contrastive loss, the previous approaches do not consider other case scenarios where there might be various instances that bear resembles to the contrastive objective. The proposal of this work focuses on adding information about the instances to relax the contrastive loss, such that fashion items do not only get their visual and textual representation closer, but similar products are not forced to be apart, resulting in a spatial structuring inspired on the fashion taxonomy.

Structured Multimodality - Catalog Organization

In this section, a model capable of creating a structured multimodal representation is presented considering two modali-

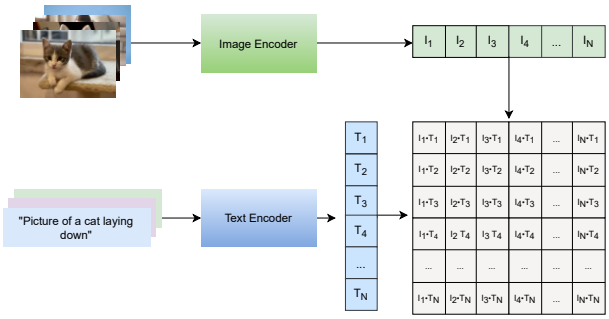

Figure 3: Training process. The image encoder is a Vision Transformer (Dosovitskiy et al. 2021); The text encoder a Transformer (Vaswani et al. 2017). Adapted from (Radford et al. 2019).

ties, visual and textual data, and prior knowledge.

Multimodal Representation

To learn a joint representation space for the two modalities, we follow a training process that has been previously explored in other works (Sohn 2016; van den Oord, Li, and Vinyals 2019; Radford et al. 2021). It consists of using mini-batches of N (image, text) pairs, corresponding to an image and a description of a product in the dataset. Each pair is processed by the model through an image and text encoders, which results in a latent vector for each modality instance, denoted by I_i and T_i, respectively. The similarity between each possible pair is then computed, using the cosine similarity, $sim(I_i, T_j)$, $i = 1, \ldots, N$, $j = 1, \ldots, N$, leading to an $N \times N$ matrix, as shown in Figure 3

Since the target pairs of instances are known, the typical approach is to minimize some variant of a contrastive loss. For instance, Radford et al. (2021) use the normalized temperature-scaled cross entropy (Chen et al. 2020), defining the loss of a positive pair (I_i, T_i) as

$$\ell(I_i, T_i) = -log \frac{\exp\left(sim(I_i, T_i)/\tau\right)}{\sum_{k=1}^{N} \exp\left(sim(I_i, T_k)/\tau\right)}, \qquad (1)$$

where τ is the temperature parameter. The global loss is then given by

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} \big(\ell(I_i, T_i) + \ell(T_i, I_i)\big). \qquad (2)$$

The drawback of this approach is that it assumes only one of the possible visual and textual embedding pairs is correct, forcing the model to approximate these and to push all other combinations farther apart. However, some of the *incorrect* pairs are often acceptable from the customer point of view, especially when the product description is not very specific. As such, the contrastive loss above is often too harsh on pairings that are not necessarily incorrect.

Prior Knowledge

In order to overcome this limitation, we propose to relax the loss function, reducing the penalization assigned to acceptable pairs. To achieve this, label relaxation is applied to the target pairs, so that the loss function is less strict with respect to which instances should paired.

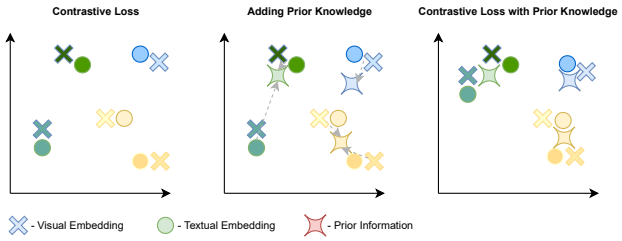

Figure 4: Adding Prior Knowledge to the Contrastive Loss.

Label Relaxation Label relaxation was originally proposed as a strategy to reduce overfitting (Lienen and Hüllermeier 2021). In this technique, the target is a set of probabilities represented in terms of an upper probability distribution, diminishing the chance of getting a biased model. The goal is to reproduce a similar form of regularization to our model such that when close instances belong to the same batch, the probability of the correct match will not be concentrated only on the exact pair, but this value is divided by the remaining elements that are not totally wrong, but are not *the* correct match.

From possibility theory (Dubois and Prade 2007), let π be the distribution that defines upper bounds on the probability of events, also called possibility distributions, such that $\pi : \mathcal{Y} \to [0, 1]$, χ is an instance space, $\mathcal{Y} = \{y_1, ..., y_k\}$ a set of class labels that, in our approach, represent the images/text intended to match, and $\mathbb{P}(\mathcal{Y})$ the space of probability distributions over χ. The intention with this approach is to keep having *the correct answer*, that is, $\pi(y_i) = 1$ for one case $y_i \in Y$. Additionally, by permitting a certain degree $\pi(y) > 0$ of plausibility to the other classes, it is possible to express that these classes are not totally wrong either through a parameter $\alpha \in [0, 1]$.

As a result, $\mathcal{Q}_\pi^\alpha$ is, at this point, given by the set of probability distributions p, that assign the probability mass of at most 1 (and at least $1 - \alpha$) to the correct pair, and at most α to the remaining acceptable pairs,

$$\mathcal{Q}_i^\alpha = \{p \in \mathbb{P}(\mathcal{Y}) \mid \sum_{y_i \neq y \in Y} p(y) \leq \alpha\} \qquad (3)$$

By introducing this label relaxation, the different views of one product and other acceptable matches are allowed to be closer in space. The objective is that when a batch with these instances is given to the model, the loss does not penalizes so much these pairings, and allowing the model to assign high probability to another view of the same product or to other pairs other than the intended match.

Figure 5 illustrates the effect the proposed label relaxation strategy has on the organization of the representation space, through the manipulation of the distribution. For illustrative purposes, a batch of N instances is considered and there are C additional pairs that are close enough to be considered appropriate for the image I_1, and being α the constant that limits the maximum probability mass attributed to these cases,

	T_1	T_2	T_3	T_4	T_5	T_6
I_1	1 - α	0	0	$\frac{\alpha}{C}$	$\frac{\alpha}{C}$	0

Figure 5: Example of the probabilities assigned to the each pair (I_1, T_i) in a batch of $N = 6$ samples, where T_4 and T_5 are two incorrect but acceptable pairs ($C = 2$).

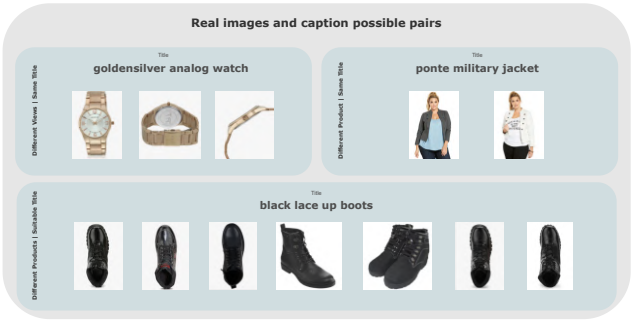

Figure 6: Different Views and Similar Instances.

Figure 5 presents the batch of assigned probabilities to those N textual instances being the pair of image I_1.

To accommodate this new target distribution, we replace the cross-entropy loss in (1) with the Kullback-Leibler (KL) Divergence, which has no computational overhead.

When using the contrastive loss with the KL divergence on batches that do not contain other acceptable pairs, the loss will be equal to the previous scenario. For the other cases, the gradients based on the optimization of the KL divergence will group attract all possible pairs, while only pushing apart unacceptable pairs.

Defining Acceptable Pairs The proposed label relaxation strategy requires the establishment of a definition of acceptable pairs. We leverage prior knowledge about the fashion domain to determine which pairs of product images and descriptions are acceptable. Specifically, the fashion taxonomy, typically used to categorize a product, allows us to determine which products share a similar root. Additionally, multiple views of a product are often part of a catalog and share a common product description. Therefore, two instances are considered close, and they should indeed be closer in the multimodal space, if they are:

1. **different views of the same product**;
2. **categorically identical**.

Examples of product images and descriptions that are considered acceptable are illustrated in Figure 6.

Experimental Setup

Dataset

In fashion, there is a wide range of datasets that focus on computer vision tasks, namely detection of multi-class labels and attributes. In multimodal retrieval, in which products are represented by images and textual descriptions, there is a lack of fashion datasets that allow accomplish this goal. The fashion domain dataset used in this work is MMD (Saha, Khapra, and Sankaranarayanan 2018), and is constituted by simulations of domain-aware in-store conversations and respective fashion products mentioned during those conversations. The dataset can be split into two relevant and different parts: the dialogues and the products. Only the latter will be used. The products are represented by different views and it is imperative that the retrieving model knows how to interpret the various points of view of an article and knows that, for instance, a photograph of the back of a dress represents the same dress as the front view, although the information extracted by the visual encoder differs. Moreover, and considering the products' categories defined by the hierarchical taxonomy, it is also possible to measure how close items are categorically. These two points are key aspects that our approach focuses on solving. Hence, the dataset is constituted by:

- Images (different views/angles), Text (titles) and Taxonomic paths (labels) from all products;
- A total of 100K fashion items, represented by 460K (image, title, taxonomy) instances.

Given the motivation of this work, there are two scenarios for the store's catalog: it can be a fixed set of products known by the system, or can be extended by new unknown instances introduced in this space, either by an **increase of the catalog**, or by the **user's queries**. For this reason, and given the lack of benchmarking datasets, the proposed solution will be assessed considering different evaluation metrics on the:

- **Static catalog** - simulation of the store's catalog. Composed by a fixed amount of instances (80% of the original data).
- **Dynamic catalog** - simulation of the store's catalog with known instances (80%) and new unknown ones (20% of the original data).

Furthermore, and to simulate the results in specific environments, special cases for the dataset and queries were also considered:

- Manipulation of the size of the catalog, aiming at testing the lack and abundance of products in a subset with 60% of the original data: $\text{MMD}_{\text{small}}$;
- Queries with different characteristics and categories to test the retrieval of unique and common products.

Evaluation Metrics

Considering that this proposal aims at structuring the multimodal space to guarantee meaningful results, the following metrics were defined to evaluate the generated catalogs:

- **Visual inspection of the retrived items** (qualitative assessment);
- The **distance between embeddings of items of the same category**;
- The **percentage of same-category products in the top K closest instances**;
- The **distance between different product views**;
- **Real Pairs: Recall@K**.

Implementation Details

Regarding the implemented model, the pre-trained CLIP (Radford et al. 2021) was used, consisting of an image and text encoders. The first is a Vision Transformer (Dosovitskiy et al. 2021) that surpasses the performance of state-of-the-art convolutional neural networks. The latter a Transformer (Vaswani et al. 2017) with a few modifications introduced in (Radford et al. 2019).

Optimizer	SGD	**learning rate** 1×10^{-4}	**momentum** 0.9
Scheduler:	OneCycleLR	**total steps** 36750	
Batch size:	750	**# Epochs**:	30

Table 1: Implemented model hyperparameters

The model was fine-tuned on the MMD dataset using SGD with momentum, with the hyperparameters defined in Table 1. The learning rate was initially set to 10^{-4} and updated with the 1 cycle policy (Smith and Topin 2019). Moreover, and with respect to the relaxation parameter that defines the upper bounds of the distribution, $\alpha = 0.25$.

Results

The following section focuses on the experiments that were conducted for evaluating the proposed solution. A qualitative assessment of the results is considered initially for the models with and without the introduction of prior knowledge. Then, the multimodal catalogs generated with the different models are evaluated in a quantitative way using the metrics previously defined.

Retrieval: Visual Evaluation

In order to evaluate the system's performance, some cases were first defined. In these tests, the aim is to understand how the model behaves when a textual, visual or multimodal query is inserted. One very important aspect to highlight is that these tests consider real scenarios, that is, the visual and textual queries do not belong to the dataset and were never seen by the architectures. Given the subjectivity of this evaluation, the results obtained are analyzed based on factors such as the category of products presented and their attributes, compared to what was expected.

Considering a query with textual information, Figure 7 represents the results obtained when performing product retrieval on $\text{MMD}_{\text{small}}$. The first row presents the instances retrieved by the architecture with the proposed relaxation, and the second the products obtained by the model with the cross entropy. It is possible to conclude that the outlier retrieved by the latter does not exists in the first's top 5.

Considering a query with textual information, Figure 8 represents the results when performing product retrieval of a non-existent product. Firstly, and regarding the instances retrieved by the architecture without prior knowledge, there are 3 items that not only do not have the intended category, but also have a very distant one (*suit* and *pajamas*). This occurrence can be justified given the existent visual similarities

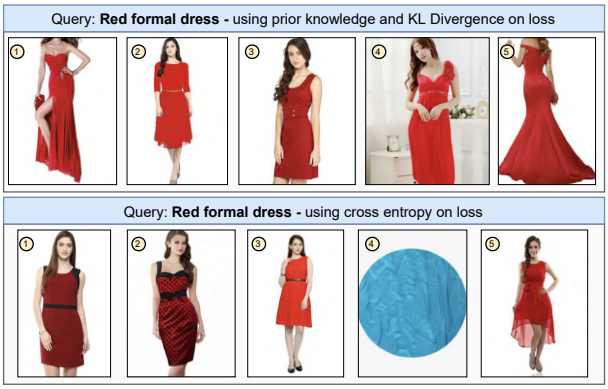

Figure 7: Top 5 closer items when trying to retrieve a "Red formal dress" with and without prior knowledge, on $\text{MMD}_{\text{small}}$.

between the first three images and a suit. On the other hand, the structure established by the proposed loss ensures that the returned products are more appropriate.

Figure 8: Top 5 closer items when trying to retrieve a "Pink suit for woman" with and without prior knowledge.

In a multimodal scenario, and with the objective of finding a product similar to an image, but with modifications that are introduced in the form of text, Figure 9 presents the outcome of the tests to the linearity of this space and the capacity of navigating in it. The figure shows the obtained results when performing both addition and removal of features from the image through textual information. Consequently, the textual embedding related to the color "*black*" has been subtracted from the embedding of the provided image (a *dress*), and the one related with the color "*red*" was summed. The results obtained after these operations are red long-sleeved dresses with the same attributes as before, but with the desired color.

Figure 10 shows the products obtained when performing both addition and removal of three adjectives in textual form from a base representation. The results achieved after these operations show the obtained contrast when retrieving opposite instances with respect to one attribute. Regarding the retrieved top, considering a visual analysis given the subjectivity of these adjectives and an evaluation as a whole,

Figure 9: Results for the top 4 closer items when trying to retrieve using an image of a long sleeve black dress, removing "*black*" and adding "*red*".

most of the results seem adequate given the baseline and subsequent modification. Even when comparing with other attributes that can be considered synonyms/antonyms, the results demonstrate consistency.

Figure 10: Results for the top 4 closer items when trying to retrieve products with a manipulated embedding.

Static Catalog

To evaluate the models in a quantitative way on the static catalog, the previously defined metrics are considered and its results presented:

1. **Distance between embeddings of items of the same category**:

   The analysis of Table 2 allows concluding that there is greater proximity between the instances of the same sub-category in the catalog generated by the model in which prior knowledge is introduced. Moreover, the smaller variance demonstrates that the increase in this proximity is in the various sub-categories existing in the dataset and not just in some.

model	mean	variance
Original Model	0.7886	0.0069
Fashion Fine-Tuned without taxonomy	0.8410	0.0034
Fashion Fine-Tuned with taxonomy	0.8760	0.0020

Table 2: Proximity (cosine-similarity) between instances of the same subcategory: mean and variance.

2. **Percentage of same-category products in the top K closest instances**:

   The neighborhood of a product is given by the instances closest to it in terms of vector similarity. By checking which is the sub-category of the products that are in the neighborhood of each instance, it is possible to evaluate if the addition of prior knowledge places products with the same sub-category in the neighborhood of each other.

   Table 3 shows that the increase in the percentage of neighbors is visible when the model is fine-tuned, but even more notorious when it is fine-tuned and prior knowledge is introduced. This increase is verified for the various values of K tested.

Top K	Original Model	Fashion Fine-Tuned without taxonomy	Fashion Fine-Tuned with taxonomy
1	0.4831	0.4873	0.4883
10	0.3531	0.3564	0.3665
25	0.3142	0.3198	0.3291
50	0.2883	0.2951	0.3013

Table 3: Percentage of same sub-category top K neighbors.

3. **Distance between different product views**:

   Items that must be close are not only those that share a common sub-category, but also the various views of a given product. Table 4 shows that for the models to which prior knowledge is added, the similarity between the views of all products is now higher. The reduction in the variance is significant when comparing the Original Model model and the Fashion Fine-Tuned with Prior Knowledge, in which it dropped to half, making the views of all products, in general, closer.

Real Pairs The previous results measure the influence of the different architectures on the spacial distancing and positioning of instances of equal sub-categories and views. The focus now is on the retrieval task, that is, measuring the influence of prior knowledge increases on the retrieval performance. Table 5 presents, for each model, the $Recall@K$ on the static catalog when matching all the 460K images with

	Original Model	Fashion Fine-Tuned without taxonomy	Fashion Fine-Tuned with taxonomy
mean	0.8604	0.9079	0.9289
std	0.1020	0.0680	0.0536
min	0.2440	0.4858	0.5820
25%	0.8168	0.8803	0.9077
50%	0.8803	0.9213	0.9399
75%	0.9331	0.9550	0.9653
max	1.0000	1.0000	1.0000

Table 4: Statistics of the cosine-similarity between all views for every product.

the respective title. The results obtained show that although the new loss emphasizes less the real pairs, the matches assigned are, for all K, more correct when prior knowledge is introduced.

Recall @K	Original Model	Fashion Fine-Tuned without taxonomy	Fashion Fine-Tuned with taxonomy
1	0.0290	0.0379	0.0401
10	0.1060	0.1501	0.1615
25	0.1631	0.2350	0.2524
50	0.2207	0.3196	0.3418

Table 5: Recall@K for $K = \{1, 10, 25, 50\}$ when retrieving the pair of an image/title on the static catalog.

model	real pairs mean probability	real pairs mean cosine-similarity
Original Model	0.1597	0.3071
Fashion Fine-Tuned without taxonomy	0.2159	0.4375
Fashion Fine-Tuned with taxonomy	0.1813	0.5300

Table 6: Probability and cosine-similarity between the real pairs from different modalities, for each model.

Regarding the similarity between embeddings calculated using the cosine-similarity, and when using CLIP and introducing prior knowledge, contrary to the trend of the probability of matching, an approximation of the embeddings that represent a match is verified. This contrast can be explained by the increase in density and spatial agglomeration of the categories, which reflects on a greater number of possible pairs when calculating all the probabilities, and a more reduced one for the real pair. In fact, this also shows that the prior knowledge is approximating same-category instances.

Dynamic Catalog: Generalization

So far, the quantitative results presented have focused on the structure of the catalog, which is static. Therefore, it was not necessary to make any considerations about the existence of never seen instances by the system, and, subsequently, its generalization capacity on a test set.

Considering the two cases in which the system receives new unknown data, two of the metrics previously used will now measure:

- **Similarity between different views, considering that a new one, unknown to the system, is introduced in the catalog**:
  Table 7 shows that when measuring the similarity between the views in the referred conditions, the proximity between them increases when the loss is regularized with prior knowledge, as it was previously seen for the different views in the static catalogue.

	Original Model	Fashion Fine-Tuned without taxonomy	Fashion Fine-Tuned with taxonomy
mean	0.8534	0.8988	0.9012
std	0.1015	0.0644	0.0636
min	0.3321	0.5390	0.5117
25%	0.8095	0.8745	0.8737
50%	0.8746	0.9130	0.9130
75%	0.9238	0.9427	0.9423
max	1.000	0.9912	0.9956

Table 7: Statistics of the cosine-similarity between the views added to the static catalog - dynamic catalog.

- **Percentage of same-category neighbors for new instances added to the static catalog**:
  This metric focuses on measuring the influence of the introduction of prior knowledge when an unknown instance is surrounded by other known ones that may share the same sub-category through the percentage of k-Nearest same-category neighbors. Table 8 shows that when instances are brought together to the baseline space, the ones created with the architecture trained with prior knowledge have higher percentage of same-category neighbors. In this case, Original Model globally outperformed the results of the Fine-Tuned model without prior knowledge, but not the one with it.

Top K	Original Model	Fashion Fine-Tuned without taxonomy	Fashion Fine-Tuned with taxonomy
1	0.5559	0.5526	0.5568
10	0.3733	0.3645	0.3686
25	0.3174	0.3081	0.3232
50	0.2904	0.2758	0.2911

Table 8: Percentage of same sub-category top K neighbors: instance from the dynamic catalog.

Real Pairs Focusing on the retrieval task, the $Recall@K$ measures the correct pairwise matches on the dynamic catalog, in which never seen instances images and captions are surrounded by other known ones when retrieving. Table 9 presents, for each model, the $Recall@K$ on the dynamic catalog when matching all the 460K images with the respective title. The results obtained show that although the new loss emphasizes less the real pairs, the matches assigned are, for all K, more correct when prior knowledge is introduced outperforming the remaining models.

Conclusions

This paper introduced a new loss function for multimodal product retrieval that takes into account prior knowledge

Recall @K	Original Model	Fashion Fine-Tuned without taxonomy	Fashion Fine-Tuned with taxonomy
1	0.0307	0.0406	0.0417
10	0.1101	0.1572	0.1643
25	0.1683	0.2446	0.2554
50	0.2272	0.3306	0.3448

Table 9: Recall@K for $K = \{1, 10, 25, 50\}$ when retrieving the pair of an image/title on the dynamic catalog.

about the fashion domain. The proposed approach learns a joint representation space for images and their corresponding textual descriptions, similarly to other contrastive-based strategies. However, the incorporation of prior knowledge works as a regularizer, alleviating the inflexibility of the popular contrastive loss by not penalizing wrong but acceptable matches between image and descriptions. The learned representation space is, thus, restructured such that products with the same taxonomical path are spatially closer. Consequently, products retrieved in this new space are less likely to be from other categories, a type of error that severely damages the quality of the results. Our results show that these improvements allowed the model to obtain better overall results, reducing the number of mistakes in which it returns products that are categorically distant and are, therefore, inappropriate. Moreover, the spatial clusters of the categories are better defined, and the views of the same product are closer together, both of which demonstrate that the system improved the structure of its representation space, leading to better performance in the product retrieval task.

References

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2017. Multimodal Machine Learning: A Survey and Taxonomy. arXiv:1705.09406.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929.

Dubois, D.; and Prade, H. 2007. Possibility theory. *Scholarpedia*, 2(10): 2074. Revision #137677.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Ranzato, M. A.; and Mikolov, T. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Gong, Y.; Ke, Q.; Isard, M.; and Lazebnik, S. 2013. A Multi-View Embedding Space for Modeling Internet Images, Tags, and their Semantics. arXiv:1212.4522.

Hironobu, Y. M.; Takahashi, H.; and Oka, R. 1999. Image-to-Word Transformation Based on Dividing and Vector Quantizing Images With Words. In *in Boltzmann machines", Neural Networks*, 405409.

Li, A.; Jabri, A.; Joulin, A.; and van der Maaten, L. 2017. Learning Visual N-Grams from Web Data. arXiv:1612.09161.

Li, S.; Lv, F.; Jin, T.; Lin, G.; Yang, K.; Zeng, X.; Wu, X.; and Ma, Q. 2021. Embedding-based Product Retrieval in Taobao Search. *CoRR*, abs/2106.09297.

Lienen, J.; and Hüllermeier, E. 2021. From Label Smoothing to Label Relaxation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10): 8583–8591.

Magalhaes, J.; Hauptmann, A. G.; Sousa, R. G.; and Santiago, C. 2021. MuCAI'21: 2nd ACM Multimedia Workshop on Multimodal Conversational AI. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5702–5703.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G. R.; Levy, R.; and Vasconcelos, N. 2010. A New Approach to Cross-Modal Multimedia Retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, 251–260. New York, NY, USA: Association for Computing Machinery. ISBN 9781605589336.

Saha, A.; Khapra, M.; and Sankaranarayanan, K. 2018. Towards Building Large Scale Multimodal Domain-Aware Conversation Systems. arXiv:1704.00200.

Smith, L. N.; and Topin, N. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, 1100612. International Society for Optics and Photonics.

Sohn, K. 2016. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

van den Oord, A.; Li, Y.; and Vinyals, O. 2019. Representation Learning with Contrastive Predictive Coding. arXiv:1807.03748.

Van Gysel, C.; de Rijke, M.; and Kanoulas, E. 2016. Learning Latent Vector Spaces for Product Search. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention Is All You Need. arXiv:1706.03762.