Neurocomputing 538 (2023) 126196

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Efficient text-image semantic search: A multi-modal vision-language approach for fashion retrieval

Gianluca Moro *[,1], Stefano Salvatori [1], Giacomo Frisoni

Department of Computer Science and Engineering – DISI, University of Bologna, Via dell'Universitá 50, Cesena (FC), Italy

ARTICLE INFO

Article history:
Received 13 July 2022
Revised 20 February 2023
Accepted 28 March 2023
Available online 31 March 2023
Communicated by Zidong Wang

Keywords:
Multi-modal retrieval
Metric learning
Vision-and-language transformers
Deep learning
Fashion domain

ABSTRACT

In this paper, we address the problem of multi-modal retrieval of fashion products. State-of-the-art (SOTA) works proposed in literature use vision-and-language transformers to assign similarity scores to joint text-image pairs, then used for sorting the results during a retrieval phase. However, this approach is inefficient since it requires coupling a query with every record in the dataset and computing a forward pass for each sample at runtime, precluding scalability to large-scale datasets. We thus propose a solution that overcomes the above limitation by combining transformers and deep metric learning to create a latent space where texts and images are separately embedded, and their spatial proximity translates into semantic similarity. Our architecture does not use convolutional neural networks to process images, allowing us to test different levels of image-processing details and metric learning losses. We vastly improve retrieval accuracy results on the FashionGen benchmark (+18.71% and +9.22% Rank@1 on Image-to-Text and Text-to-Image, respectively) while being up to 512x faster. Finally, we analyze the speed-up obtainable by different approximate nearest neighbor retrieval strategies—an optimization precluded to current SOTA contributions. We release our solution as a web application available at https://disi-unibo-nlp.github.io/projects/fashion_retrieval/.

1. Introduction

Text and image multi-modal retrieval is the problem of using a query represented by a sentence or an image to retrieve other texts or images in a dataset. Tackling this task in the fashion domain is notoriously challenging [1]. Specifically, it demands the ability to manage queries referring to fine-grained details of clothes within an online catalog, comprising attributes and noun phrases (e.g., "*Low sneakers in black polished leather, round toe, closure with tone-on-tone laces, padded tongue and collar.*"). Meeting this requirement is indeed not requested by the retrieval in the general domain where, on the contrary, the focus is on coarse-grained objects inside the image, and descriptions are mainly high-level, like "*a little dog is holding a ball in his mouth*" or "*a yellow fire hydrant in front of a blue wall*".

State-of-the-art (SOTA) results for fashion retrieval have been recently achieved by vision-and-language (V + L) transformers, i.e., transformers [2] able to process both texts and images in a single architecture. Despite the remarkable achievements, they suffer from a major drawback: input modalities still have to be strictly coupled and jointly processed to produce a similarity score that is used to get the most relevant results to a query. This means that a forward step must be computed for each query-document pair at test time, which is a highly inefficient solution if one must perform retrieval on large-scale datasets.

In this paper, we propose integrating deep metric learning (DML) after self-supervised pretraining to boost multi-modal retrieval efficiency and efficacy, with the same model usable as an image and/or text encoder. Mechanically, our architecture foresees a single V + L transformer with two-stage training. Firstly, texts and images interact within the model, learning deep relationships between them. Then, learned weights are decoupled, and DML is used to construct a latent space in which texts and images are separately embedded, where the distance between points translates into a measure of similarity between the actual data. We demonstrate that deep multi-modal relationships acquired via self-supervision can be exploited by DML to tune text and image representations further. Note that this approach differs from existing solutions, which either utilize a single interaction-focused V + L transformer or two different text and image encoders trained with DML and a representation-focused architecture.

* Corresponding author.
E-mail addresses: gianluca.moro@unibo.it (G. Moro), s.salvatori@unibo.it (S. Salvatori), giacomo.frisoni@unibo.it (G. Frisoni).
[1] Equal Contribution.

https://doi.org/10.1016/j.neucom.2023.03.057
0925-2312/

By decoupling text and image embeddings, we take only one inference step at runtime to encode the query, which can then be compared with document embeddings pre-computed offline. This design also enables the adoption of multidimensional indices with which multi-modal retrieval is significantly faster. Retrieval performance similitude across different datasets confirms our solution's goodness and ability to generalize.

We conduct extensive experiments by testing different ranking losses in the DML phase, showing how they compare with the triplet loss, which is still the most used loss function in the multi-modal retrieval literature. Unlike previous contributions, our architecture fully replies to the real-world needs imposed by fashion product retrieval, non requiring each image to be processed with pretrained convolutional neural networks (CNNs) and capturing fine-grained details.

Our transformer architecture with a two-stage training approach pushes SOTA Rank@*K* accuracy for Text-to-Image and Image-to-Text retrieval on the well-known FashionGen [3] dataset, also registering promising results on DeepFashion-Synthesis [4]. Additionally, we show that our model is much more time-efficient, up to 512x faster than the SOTA ones, and could be used on large-scale datasets using multi-dimensional indices and approximate nearest neighbor semantic search to speed up retrieval.

2. Related Work

2.1. Transformer-based Fashion Multi-Modal Retrieval

Multi-modal retrieval in the fashion domain has been recently addressed by FashionBERT [1] and KaleidoBERT [5] V + L transformers. Contrary to the latest contributions in general-domain retrieval with transformers [6–10], such architectures do not rely on extracting regions of interest (RoIs), which notoriously tend to ignore small-grained details. Instead, images are subdivided into square patches and processed via a pretrained ResNet [11] model. However, operating with ResNet translates into storing a pre-fixed 2048D vector for each patch which (i) produces a negative impact in terms of memory and time complexity, (ii) does not scale to smaller image patches or images with higher resolution. Motivated by these issues, some up-to-date models—such as MVLT [12]—directly manage the raw visual patches without extra frozen ResNet preprocessing models and introduce a masked image reconstruction task (end-to-end pretraining scheme). Nevertheless, they ignore DML.

Moreover, traditional transformer-based solutions necessitate texts and images to be coupled and simultaneously fed to the model, which outputs a similarity score between them. As addressed in [13–16], this excludes the possibility of defining multi-dimensional indices on embeddings (e.g., Inverted File Indices, PCA dimensionality reduction, Product Quantization [17]), thereby precluding applications on massive datasets where fast retrieval methods are essential. Indeed, the necessity of traversing every query and gallery item pair in single-stream architectures frequently causes unacceptable speed in cross-modal retrieval applications. Although the authors of FashionBERT propose a masking strategy for using their model with the least waste of time and effort, they do not provide any supporting experiments; indeed, we find it not working well in practice (see D). For these reasons, the community is witnessing a transition toward dual encoding.

We introduce a V + L transformer that does not use ResNet to process patches. This leads to a *lighter architecture* that allows experimenting with *different levels of image-processing details* easily. In our work, we also show that it is possible to combine V + L transformers and DML through a two-stage training process designed to generate *cross-modal-aware individual latent representations of texts and images* that we prove to be more efficient in large-scale retrieval scenarios.

2.2. Multi-Modal Retrieval with Deep Metric Learning

Deep metric learning has caught the attention of many researchers dedicated to natural language processing tasks [18,19], with non-negligible effectiveness also established in low-resource regimes. An efficient solution for multi-modal retrieval has been proposed by Frome et al. in VSE [20]; a skip-gram model and a CNN are trained with triplet loss [21] to generate a common text-image latent space where semantic similarity is measured with the Euclidean distance. Subsequently, several contributions emerged to enhance its efficacy. For instance, VSE++ [22] added the hard mining technique during the training phase to focus on more informative samples. SCAN [23] and PFAN [24] later suggested including some interaction between text and image representations in the form of ad hoc attention layers; these works also differ from VSE and VSE++ since they first introduced Fast R-CNNs [25] to extract RoIs from images and help the model concentrate on the critical parts of an image. Lately, CLIP [26], and in particular its fashion adaptation FashionCLIP [27], proposed to use contrastive pretraining with separate image and text encoders to generate a shared latent space in which performing several multi-modal tasks. Belonging to the same family, Shin et al. proposed e-CLIP [28], the result of the industrial application of CLIP on a massive e-commerce dataset consisting of 330 M pairs (no significant architecture innovations). CMA-CLIP [29] extended a pre-trained CLIP with two types of cross-modality attention, i.e., sequence-wise and modality-wise, the first designed to model fine-grained patch-token relationships, the second to weigh each input modality by its relevance for the downstream task. Yet, CMA-CLIP and e-CLIP are not tested on retrieval tasks, with the latter not even assessed on FashionGen.

After our first submission, FaD-VLP [14] and FashionViL [15] exhibited significant gains on FashionGen, utilizing contrastive learning to align text and image representations. FaD-VLP is a decoder-based model architecture empowered by fashion-specific pretraining, including two tasks centered on weakly-supervised triplets. On the other side, FashionViL consists of three encoders (text, image, text + image) and two V + L pretraining tasks leveraging fashion data specialties, i.e., multiple images/views for the same product and attribute-rich descriptions. Differently from all previous works, FaD-VLP and FashionViL are the only models to trade on multiple heterogeneous datasets for pretraining (other than FashionGen), being not directly comparable for fairness and reaching more than 6× our instance count. Furthermore, they both use ResNet50 for image encoding and—as explicitly declared in the original publications—are resource-hungry, with FaD-VLP relying on two 8 GPU NVIDIA A100 nodes for pretraining.

Another concurrent work published during the revision of this paper is CommerceMM [16]. It grasps text-image pairwise relations through contrastive learning and nine new pretraining tasks on cross-modal and cross-pair data. By contra, it counts a single pretraining phase, relies on trainable image encoders (e.g., ResNet50/ViT), and is not pretrained on FashionGen.

Our contributions are not focused on designing novel pretraining tasks but on exploring an original and highly-efficient combination of multi-modal self-supervision and DML. In contrast to the aforementioned solutions, we (i) employ a *ResNet-free V + L self-supervised training* with a multi-task loss to uncover hidden semantic relationships between texts and images, (ii) test *multiple DML losses* that, as we demonstrate, provide better results com-

pared to the widely used triplet loss, (iii) achieve new state-of-the-art performance, keeping the same pretraining dataset.

3. Methodology

3.1. Model Architecture

Fig. 1 depicts our model's architecture. We design a BERT-like transformer with additional layers to process visual and textual inputs. Since we apply our model to a multi-modal retrieval task in the fashion domain, texts and images are, in our case, descriptions and photos of fashion products.

Each text is tokenized using the SentencePiece vocabulary borrowed from RoBERTa [30]. Two special tokens (`[CLS]` and `[SEP]`) are attached at the beginning and end of the sequence. Each token is processed through a text embedding layer that transforms it into a dense vector $t \in \mathbb{R}^{768}$. Positional embeddings are then added to the representation thus obtained. Finally, segmentation embeddings are employed to differentiate it from visual input: we use an array of zeros for texts and an array of ones for images.

We split each image into fixed-size square patches that are then flattened in a linear sequence: an $H \times W$ image with C channels $i \in \mathbb{R}^{W \times H \times C}$ is transformed into a one-dimensional tensor $s \in \mathbb{R}^{N \times (P^2 \cdot C)}$ where P is the patch size and $N = \frac{HW}{P^2}$ is the total number of patches. Each patch is projected into the same $\mathbb{R}^{768}$ space of text tokens through a linear projection layer. Positional and segmentation embeddings are also added to the resulting vector. Text and image representations are then concatenated $[t; s]$ and fed into a transformer encoder, where they can interact through the self-attention layers.

We do not extract RoIs from images, as it has already been proven ineffective in the fashion domain [1]. Unlike FashionBERT and KaleidoBERT, however, our solution doesn't even depend on pretrained CNN networks (e.g., ResNet) to process images, and each patch is projected through a single linear layer. This allows us to effortlessly choose the granularity with which images are subdivided and select the best detail level to work with—a practice otherwise heavily computationally expensive (e.g., halving the patch size would result in 4 times the number of patches and an equivalent number of ResNet forward steps).

3.2. Pretraining

We pretrain the model on text-image pairs with a multi-task loss to learn semantic relationships between visual and textual inputs. The training tasks are described in the following paragraphs.

Text-Image Alignment (TIA): We use the output of the `[CLS]` token as input to a binary classifier that must predict whether the given text and its paired image are related to each other (in our case "related" means that the text is describing the image). Given a text-image pair (t, p) sampled from the dataset D and the score $s(t, p)$ returned by the classifier, we use the binary cross-entropy loss (Eq. 1).

$$L_{TIA}(\theta) = -E_{(t,p)\sim D}[y \log s_\theta(t,p) + (1-y)\log(1 - s_\theta(t,p))], \tag{1}$$

where y is the true label of the input and θ denotes the model weights that can change during training.

Masked Language Modeling (MLM): We randomly mask 15% of input text tokens: 80% of the time they are replaced with a special `[MASK]` token, 10% with a random word, and 10% they stay unchanged. Denoting with $t_{\setminus i} = \{t_1, t_2, \ldots, [\texttt{MASK}]_i, \ldots, t_n\}$ the input sentence in which the i-th token has been masked, the network must minimize the loss depicted in Eq. 2 for all masked tokens t_i.

$$L_{MLM}(\theta) = -E_{(t,p)\sim D}[\log P_\theta(t_i | t_{\setminus i}, p)], \tag{2}$$

where $P_\theta(t_i | t_{\setminus i}, p)$ signifies the probability assigned to the masked-out token t_i by the model given its surrounding text $t_{\setminus i}$ and image patches p.

Masked Patch Prediction (MPP): We randomly mask images instead of descriptions, corrupting 20% of image patches as fol-

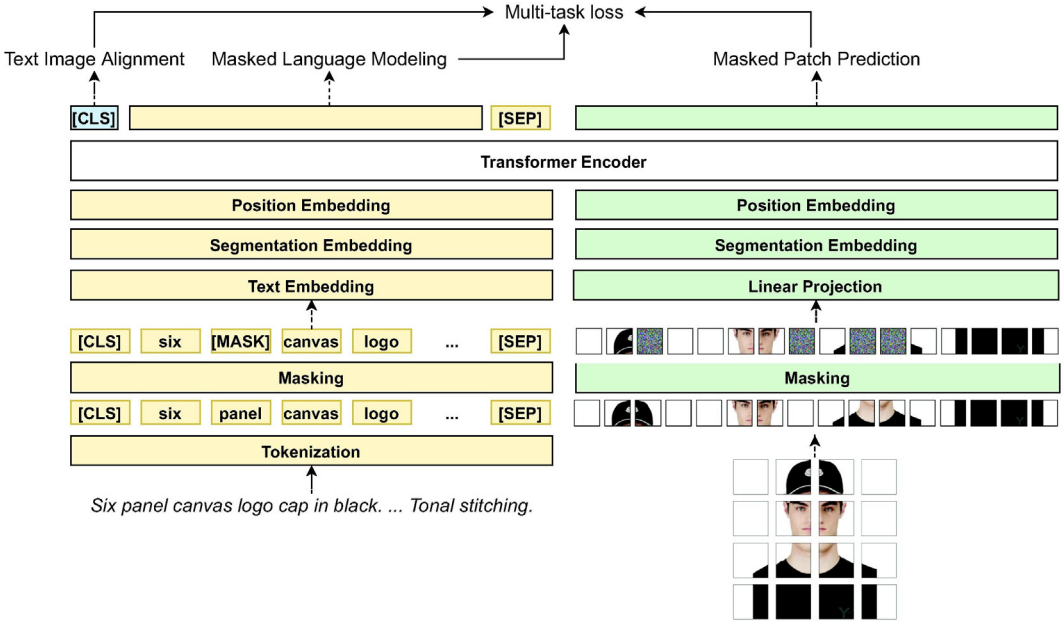


Fig. 1. Model architecture. On the language side, descriptions are processed following the BERT procedure: they are tokenized and masked with a fixed probability, then an intermediate representation is extracted for each token using text, segmentation, and position embedding layers. On the vision side, images get split into squared patches that—similarly to text processing—are masked and processed through image, segmentation, and position embedding layers. The representations are then concatenated and fed into a multi-modal fusion encoder (transformer-based). The model is trained on three tasks: Text-Image Alignment, Masked Language Modeling, and Masked Patch Prediction.

lows: 80% of the time, we substitute them with a random one, 10% of the time, we replace them with another patch in the image; we leave the remaining 10% untouched. As proposed in [31], we ask the model to predict the 3-bit mean color of each masked patch, thus having a classification problem with $(2^3)^C$ classes (=512 in our case) where C is the number of channels. Given the input patches $p = \{p_1, p_2, \ldots, p_m\}$, we denote with $p_{\setminus i} = \{p_1, p_2, \ldots, \texttt{[MASK]}_i, \ldots, p_m\}$ the sequence in which the i-th patch has been masked. For every masked patch, we minimize a cross-entropy loss similar to the MLM one (see Eq. 3).

$$L_{MPP}(\theta) = -E_{(t,p)\sim D}[\log P_\theta(\bar{p}_i | p_{\setminus i}, t)], \tag{3}$$

where $P_\theta(\bar{p}_i | p_{\setminus i}, t)$ marks the probability assigned by the model to the target 3-bit mean color $\bar{p}_i$ conditioned on the surrounding patches $p_{\setminus i}$ and the input description t.

The final loss—shown in Eq. 4—is the sum of the three losses.

$$L(\theta) = L_{TIA}(\theta) + L_{MLM}(\theta) + L_{MPP}(\theta). \tag{4}$$

3.3. Metric Learning

After completing the first training phase (Section 3.2), we perform a second training using the obtained weights to learn a latent space where both texts and images can be individually embedded. We point out this is not immediately possible since the model requires both a text span and an image in input for early fusion. To overcome this limit, we leverage our two distinct encoding channels (Fig. 1). Pointedly, when a modality is provided in input, we turn off the encoding channel related to the other modality (Fig. 2). In both cases, we take the embedding $v \in \mathbb{R}^{768}$ from the `[CLS]` token as a latent representation for the given input and use it to train the model with a metric learning loss. By doing so, we merge one-stream and two-stream architectures to complement each other's inadequacies, having in the second stage a model that can work both as a text encoder and an image encoder.

Previous works that have tackled multi-modal retrieval with DML have primarily used triplet loss to train their models [20,22,23]. In our work, we also analyze other losses proposed more recently in the literature, namely: Angular Loss [32] and Multi-similarity Loss [33].

In *Triplet Loss*, three elements (i.e., anchor, positive, and negative) are combined to form a triplet of embeddings (x_a, x_p, x_n) such that $y_a = y_p$ and $y_a, y_p \neq y_n$, where y denotes the label of a given element. The goal is for the distance between the anchor and the negative to be greater than the distance between the anchor and the positive, at least by a margin m, which is left as a hyperparameter. Given a batch $\mathcal{B}$ of N samples, the formula for the loss is reported in Eq. 5.

$$L = \frac{1}{N}\sum_{i=1}^{N} \max(0, d(x_a^{(i)}, x_p^{(i)}) - d(x_a^{(i)} - x_n^{(i)}) + m), \tag{5}$$

d is a distance function (e.g., euclidean, cosine).

The *Angular Loss* was presented to improve on the classical Triplet Loss by not considering the distance between anchor, positive and negative samples but focusing instead on the angle between them (Eq. 6).

$$L = \frac{1}{N}\sum_{x_a \in \mathcal{B}} \left\{ \log \left[1 + \sum_{x_n \in \mathcal{B}} \exp(f_{a,p,n}) \right] \right\}, \tag{6}$$

with $f_{a,p,n} = 4\tan^2\alpha(x_a + x_p)^\top x_n - 2(1+\tan^2\alpha)x_a^\top x_p$ and α hyperparameter.

The *Multi-similarity Loss* is the most recent of the three. It is based on the assumption that multiple similarities between anchors, positives, and negatives should be considered when sampling and weighting pairs so that the most informative ones are used during training. The corresponding loss is formulated in Eq. 7.

$$L = \frac{1}{N}\sum_{i=1}^{N} \left\{ \frac{1}{\alpha} \log \left[1 + \sum_{k \in \mathcal{P}_i} e^{-\alpha(S_{ik} - \lambda)} \right] + \frac{1}{\beta} \log \left[1 + \sum_{k \in \mathcal{N}_i} e^{-\beta(S_{ik} - \lambda)} \right] \right\}, \tag{7}$$

where S_{ij} represents the similarity between samples i and j (e.g., cosine similarity or dot product); α, β and λ are hyperparameters, $\mathcal{P}_i$ and $\mathcal{N}_i$ are the index sets of positive and negative pairs for the anchor x_i defined as $\mathcal{N}_i = \left\{ k | S_{ik} > \min_{y_i = y_j} S_{ij} - \epsilon \right\}$ and $\mathcal{P}_i = \left\{ k | S_{ik} < \max_{y_i \neq y_j} S_{ij} + \epsilon \right\}$, where ϵ is a hyperparameter.

Once the model is trained with one of the aforementioned ranking losses, we can perform multi-modal retrieval (text-to-image or image-to-text) in the following way. Given a query q (text or image) and a list of candidate documents (texts or images) $D = \{d_1, \ldots, d_n\}$:

1. the query q is fed into the model to obtain the corresponding embedding vector;
2. the process is repeated for all $d \in D$ (note that this operation should be performed offline to improve efficiency);

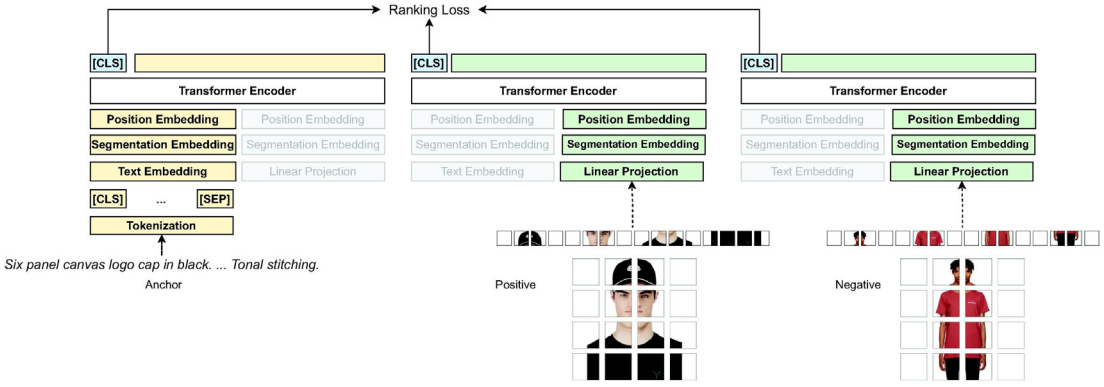


Fig. 2. Training architecture for metric learning. When a text is given in input, the image processing channel is disabled; in the same way, if an image is provided, the text processing channel is not utilized. The output vector relative to the `CLS` token acts as a latent representation for both modalities. Given a description, we take the image from the same product or a different one to construct positive and negative pairs, respectively. The figure portrays an example where the description is used as an anchor element, the corresponding image as a positive element, and another random image as a negative sample. This order is not mandatory, and both texts and images are used as anchors, positive or negative elements during our training.

3. the nearest neighbor search (k-NN) is conducted between the query embedding and document embeddings (this can be optimized using indexing strategies);
4. the nearest documents are the ones considered to be more relevant to the given query.

A summary of this procedure applied to a text-to-image retrieval scenario is illustrated in Fig. 3.

4. Experiments

4.1. Dataset

We test our model on two distinct datasets: FashionGen [3] and DeepFashion-Synthesis [4]. FashionGen contains 293,008 images (256x256 size) of fashion products paired with textual descriptions provided by professional stylists. It includes a total of 67,666 products photographed several times at different angles up to a maximum of 6. Besides a textual description, each product is also accompanied by a category (e.g., Tops, Pants, Boots) and a subcategory (e.g., Wingtip Boots, Desert Boots, Tall Boots). The dataset is split up into 260,480 records for training and 32,528 for validation.

DeepFashion-Synthesis [4] comprises a subset of 78,979 images (128x128 size) from the DeepFashion attribute dataset, in which the person faces the camera and the image's background is not too noisy. Each photo has one brief sentence expressing the clothes' visual characteristics (e.g., the color, texture, or the sleeves' length). 70000 images are used for training and the remaining 8979 for validation.

For pretraining, we extract two records for each entry in the dataset: one positive $< text, image >$ pair in which we take the description of a product with its corresponding image, and one negative $< text, image >$ pair in which the same description is paired with a random image taken from the same subcategory. For FashionGen, we end up with a total of 520,960 pairs for training and 65,056 for validation; for DeepFashion-Synthesis, we have 140,000 and 17,958 records for training and validation, respectively.

For DML, we use an online triplet mining strategy at each iteration: we take a batch of N random products and encode their description and image, producing a total of $2N$ embeddings. Each of them is used once as an anchor element considering the description or the image coming from the same product as a valid positive sample; the remaining text and image embeddings extracted from different products are used as negative elements.

Please note that both datasets are not designed for retrieval tasks. Given the unavailability of baseline text-to-image and image-to-text retrieval scores for DeepFashion-Synthesis and the impossibility of faithfully generating image features with FashionBERT on other datasets than those considered by the authors[2], we conduct main head-to-head comparisons and ablation studies on FashionGen. Notably, DeepFashion-Synthesis further enriches our evaluation setup by capturing diverse multi-modal settings and challenges. In fact, contrasted with FashionGen, it is characterized by "simpler" images of everyday clothes, which are described in less detail with shorter captions.

4.2. Results

We evaluate our model on two multi-modal retrieval tasks: Text-to-Image and Image-to-Text retrieval. For fair comparisons, we adopt the same evaluation methods used in [1,5]. *All the models are evaluated under the same conditions, exclusively exploiting FashionGen*[3]. Details about models, training, and hardware configurations are listed in A and B.

1. **Text-to-Image Retrieval (TIR)**: Given a product description in the dataset, the model is asked to find the corresponding image among 100 other random images of products from the same subcategory.
2. **Image-to-Text Retrieval (ITR)**: Given an image of a product in the dataset, the model is asked to find the corresponding description among 100 other random descriptions of products from the same subcategory.

First, we run some experiments on the FashionGen dataset, testing multiple models that differ by the patch size used to split images and the ranking loss employed. In particular, we test 32x32 and 16x16 patch sizes (an additional examination with 8x8 patches is reported in Section 5). Model performances are evaluated using the Rank@K metric (with $K = \{1, 5, 10\}$) that measures how many times the correct image or text appears in the first K retrieved documents. The results are reported in Table 1.

Multi-similarity loss obtains the highest scores compared to the Triplet and Angular losses. We think this is due to its weighting and mining strategy, allowing the model to train on the most informative samples. We also observe that a fine-grained subdivision with smaller image patches leads to higher accuracy for all ranking losses, proving how greater detail allows the model to understand more complex relationships between an image and its description. The best model overall is the one that uses 16x16 patches and is trained with Multi-similarity loss.

In Table 2 we compare our results with other contributions proposed in literature tested on the FashionGen dataset, in particular: VSE [20] and VSE++ [22] project texts and images into a joint latent space using an LSTM as text encoder and ResNet as image encoder, both trained with Triplet Loss; SCAN [23] and PFAN [24] add RoIs extraction and ad hoc attention mechanisms (but without employing the full transformer architecture); ImageBERT [6], ViLBERT [10], VLBERT [7] and OSCAR [9] are vision-and-language transformers that process RoIs and text tokens in a single architecture that is trained to assign a similarity score to text-image pairs; FashionBERT [1] and Kaleido-BERT [5] employ a similar strategy but they work by subdividing images into squared patches that are then transformed through ResNet into dense vectors that are processed together with text tokens (they are the current SOTA transformer models for multi-modal retrieval in the fashion domain).

The results shown in Table 2 demonstrate that we outperform all previous works proposed in the literature, corroborating the effectiveness of our solution even in the scenario with non-interacting text and image embeddings inside the self-attention layers. Moreover, we conducted an experiment training our model on a second dataset, Deepfashion-synthesis, to prove that our approach also works well on other sets of images and texts. The results are showcased in Table 3 compared to the ones obtained on FashionGen. In general, the accuracies are similar for the two datasets, despite the differences in the type of texts and images available. The descriptions of DeepFashion-Synthesis are, in fact, shorter on average than the ones in FashionGen and, therefore, less specific. However, we do not have a notion of 'subcategory' in the DeepFashion dataset, so the candidate texts and images for each query are sampled randomly from the whole test set, making the

[2] FashionBERT unknown image preprocessing pipelines and ResNet implementation inconsistencies (ResNet-50 instead of ResNeXt-101), https://github.com/alibaba/EasyTransfer/issues/28.

[3] We exclude [14–16,28,29,12] from our baselines due to different pretraining datasets or code unavailability (i.e., reproducibility and equal-settings judgment not possible).

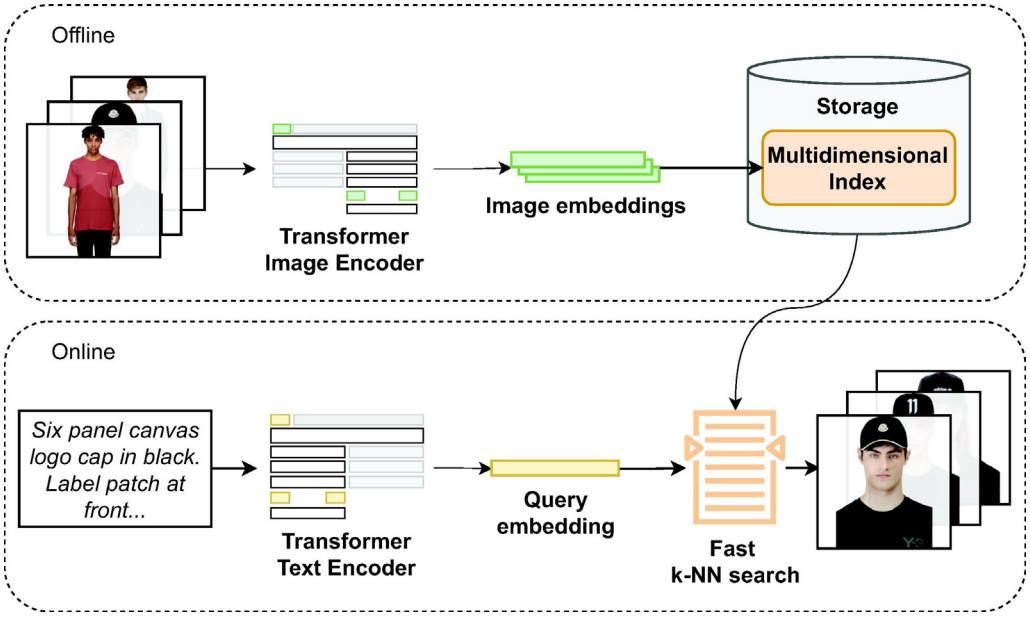


Fig. 3. Example of text-to-image retrieval with our architecture. The top section shows the operations to be performed offline to store image embeddings and organize them using a multi-dimensional index. The bottom part shows the real-time retrieval of relevant results starting from the query provided by the user.

Table 1
Results obtained by training our model using different patch sizes (P = 32, 16) and different ranking losses (MS = Multi-similarity, AN = Angular, TP = Triplet) on Image-to-Text and Text-to-Image retrieval tasks on the FashionGen dataset. The best results are highlighted in bold.

Model	TIR			ITR		
	Rank@1	Rank@5	Rank@10	Rank@1	Rank@5	Rank@10
Patch Size = 32 + MS	39.5%	72.5%	84.5%	38.5%	75.4%	85.6%
Patch Size = 16 + MS	**43.1%**	**76.6%**	**87.6%**	**46.7%**	**80.0%**	**89.3%**
Patch Size = 32 + AN	19.8%	49.6%	64.4%	20.8%	55.9%	69.7%
Patch Size = 16 + AN	21.8%	52.2%	68.8%	26.2%	57.7%	73.1%
Patch Size = 32 + TP	21.5%	52.2%	68.2%	22.5%	54.5%	70.3%
Patch Size = 16 + TP	24.1%	55.7%	69.9%	21.2%	55.0%	70.8%

Table 2
Results obtained by our best model (Multi-similarity loss with 16x16 patches) compared to previous solutions proposed in the literature for Text-to-Image (TIR) and Image-to-Text (ITR) retrieval on the FashionGen dataset. We also report the *SumR* metric for each model, computed as Rank@1 + Rank@5 + Rank@10. The best results are highlighted in bold.

Tasks		**VSE**	**VSE++**	**SCAN**	**PFAN**	**ViLBERT**	**VLBERT**	**Image Bert**	**OSCAR**	**Fashion Bert**	**Kaleido Bert**	**Our P = 16 + MS**
ITR	R@1	4.01%	4.59%	4.59%	4.29%	20.97%	19.26%	22.76%	23.39%	23.96%	27.99%	**46.70%**
	R@5	11.03%	14.99%	16.50%	14.90%	40.49%	39.90%	41.89%	44.67%	46.31%	60.09%	**80.00%**
	R@10	22.14%	24.10%	26.60%	24.20%	48.21%	46.05%	50.77%	52.55%	52.12%	68.37%	**89.30%**
TIR	R@1	4.35%	4.60%	4.30%	6.20%	21.12%	22.63%	24.78%	25.10%	26.75%	33.88%	**43.10%**
	R@5	12.76%	16.89%	13.00%	20.79%	37.23%	36.48%	45.20%	49.14%	46.48%	60.60%	**76.60%**
	R@10	20.91%	28.99%	22.30%	31.52%	50.11%	48.52%	55.90%	56.68%	55.74%	68.59%	**87.60%**
SumR		75.20	94.16	87.29	101.90	218.13	212.84	251.36	241.30	251.53	319.52	**423.30**

Table 3
Results obtained by our best model (patch size P = 16, multi-similarity ranking loss MS) on the DeepFashion-Synthesis dataset compared to FashionGen.

Dataset	TIR			ITR			*SumR*
	Rank@1	Rank@5	Rank@10	Rank@1	Rank@5	Rank@10	
FashionGen	43.10%	76.60%	87.60%	46.70%	80.00%	89.30%	423.30
DeepFashion-Synthesis	41.40%	82.20%	92.50%	48.30%	82.35%	92.70%	439.45

task less challenging for our model. Nonetheless, the results in Table 3 demonstrate that our architecture generally works with long and short captions and different-resolution images.

Moreover, we argue that our solution is efficient since document embeddings can be pre-computed, and the nearest neighbor search can be optimized using multi-dimensional indexing strategies. We further examine this claim in Section 4.3. For qualitative evaluations (i.e., t-SNE visualization and retrieval examples) and a reference guide for our open web application, the reader is referred to C and E.

4.3. Retrieval Efficiency

One of the advantages of decoupling text and image embeddings is that it is possible to index the vectors offline and perform retrieval by nearest neighbor search significantly faster. In this section, we quantify the efficiency gain by measuring the time required to retrieve the top 100 documents relative to a given query on different database sizes. We compare the times required by our model with FashionBERT to show the advantages of our solution compared to the state-of-the-art. We test 3 different strategies for our model: (i) Naïve k-NN, which stores full vectors in memory and computes the exact distance between them; (ii) Inverted File Index (IVF) strategy, which segments the dataset into a fixed number of Voronoi cells in the multi-dimensional space and, at search time, only compares the query vector with the vectors contained in the same cell; (iii) Inverted File Index with PCA reduction, which performs IVF retrieval after reducing the dimension of the vectors to a fixed configurable number. The last two tests show that having separate text and image embeddings makes it possible to use approximate nearest-neighbor techniques that are much more efficient on large datasets. Fig. 4 shows the time required by each index type to retrieve the top 100 documents on samples of 10.000, 100.000, and 1.000.000 records.

As we see, our solution is up to 512x faster than FashionBERT. Moreover, IVF and PCA reduction lead to significantly higher performances with respect to the Naïve Search, especially for large database sizes. Previous SOTA approaches, such as FashionBERT and Kaleido-BERT, cannot benefit from these types of efficient indexing strategies, and their response time exceeds 2 s already with a dataset of 10.000 records.

5. Ablation Studies

5.1. Pretraining and Metric Learning Contribution

To test the contribution of the different training phases, we conducted the following ablation studies on the FashionGen dataset: (i) we took the pretrained model without metric learning and used the Text Image Alignment score to reorder the retrieval results; (ii) we ran an experiment training the model with metric learning only, skipping the pretraining phase; (iii) we implemented a solution similar to VSE which, however, uses BERT as text encoder, ResNet as image encoder, and trains them with multi-similarity loss. The results of the experiments are reported in Table 4 (with our best model for comparison).

Our solution, which combines both pretraining and metric learning, is the one that obtains the highest accuracy. The models that use only one of the two phases bring worse results, proving that the combination of self-supervised pretraining with transformers and DML leads to better multi-modal retrieval capabilities. Despite getting lower metrics, the other two models are still slightly better than KaleidoBERT; this also means that the transformer architecture we propose, with the subdivision of images in more fine-grained patches and the removal of ResNet preprocessing, contributes to improving the performance of the model.

Regarding the BERT + ResNet model, we can see that it achieves rather good results; this confirms once again the effectiveness of the multi-similarity loss, used in this case with two separate pretrained models such as BERT and ResNet. However, we note that our solution is still more effective, proving the advantages of our pretraining phase in which texts and images interact within the attention layers.

5.2. Fine-grained subdivision

Results from Table 1 show that the models with a patch size of 16 pixels perform better than the ones with 32x32 patches. This is probably because using a fine-grained subdivision leads the model to focus on smaller details that provide a better understanding of the image. We have experimented with FashionGen, training the model using 8x8 patches to test this claim further.

However, one drawback of reducing the patch size is that the resulting number of patches increases quadratically. Setting the size to 8 pixels leads indeed to a total of 1024 patches for each image (and a longer sequence length overall if we consider text tokens). Using quadratic attention, pretraining would have ended in ∼320 h (almost two weeks), and batch size should have been lowered to 2 records. To reduce training time and keep the same batch size used in the other experiments, we have thus replaced the classic quadratic attention with Performer [34] linear attention—popular to manage long inputs without operating chunk-level segmentation [35]—reducing training time from ∼320 h to ∼175 h and keeping the same batch size. We only tested multi-similarity loss in the metric learning phase since it was the one that led to the best results in the previous experiments. Results are reported in Table 5.

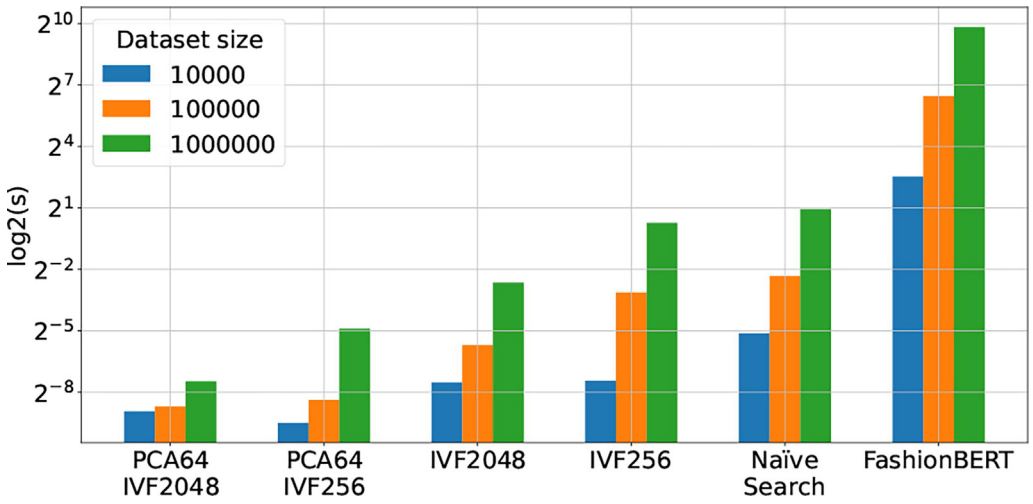


Fig. 4. Time required by different retrieval strategies to process 1000 queries on databases of different sizes. `IVF<X>` is the inverted file index that subdivides the space into `X` cells; `PCA64` at the beginning indicates that the vectors are first reduced with PCA from $\mathbb{R}^{768}$ to $\mathbb{R}^{64}$.

Table 4
Contribution of the different training phases of our architecture in downstream results. The best results are highlighted in bold.

Model	TIR			ITR			*SumR*
	Rank@1	Rank@5	Rank@10	Rank@1	Rank@5	Rank@10	
Pretrain Only	36.3%	71.7%	87.2%	37.0%	75.3%	89.0%	396.5
Metric Learning Only	33.6%	62.6%	76.2%	36.2%	64.2%	77.6%	350.4
BERT + ResNet	41.8%	72.5%	81.6%	44.2%	79.9%	88.9%	408.9
Pretrain + Metric Learning (Our)	**43.1%**	**76.6%**	**87.6%**	**46.7%**	**80.0%**	**89.3%**	**423.3**

Table 5
Results obtained training performer on FashionGen with 8x8 patches. We report for comparison also the results of the models with a patch size of 16x16 and 32x32 pixels. The best results are highlighted in bold.

Model	TIR			ITR			*SumR*
	Rank@1	Rank@5	Rank@10	Rank@1	Rank@5	Rank@10	
Patch Size = 8 (Performer)	41.8%	73.7%	87.1%	45.7%	77.8%	87.8%	413.9
Patch Size = 16	**43.1%**	**76.6%**	**87.6%**	**46.7%**	**80.0%**	**89.3%**	**423.3**
Patch Size = 32	39.5%	72.5%	84.5%	38.5%	75.4%	85.6%	396.0

While there is an improvement going from 32x32 patches to 16x16, we can see that this does not happen when going from 16x16 to 8x8 patches. One reason may be that this granularity is too fine-grained and does not allow the model to focus on truly relevant parts of the images. Another possibility could be that the approximation introduced by Performer to optimize the attention mechanism is not accurate enough, leading to lower accuracy.

6. Conclusions

We presented an efficient and effective approach for Text-to-Image and Image-to-Text retrieval in the fashion domain that combines vision-and-language transformers and deep metric learning. In contrast with SOTA solutions, where texts and images are coupled and simultaneously fed to the model to produce a similarity score, we propose to decouple text and image embeddings through a two-stage training phase. Deep metric learning generates a latent space where texts and images can be individually embedded. Their distance translates into a semantic similarity score that can be used to sort retrieval results. We proved the efficacy of our solution in the fashion domain, improving all SOTA Rank@*K* accuracies for Text-to-Image and Image-to-Text retrieval on the FashionGen dataset. In addition, we conducted several ablation studies testing different ranking losses for metric learning and different levels of detail (i.e., number of patches) with which images are fed into the model, showing how the combination of these influences the overall results. Lastly, we run experiments to prove the efficiency of our model on large-scale datasets with and without indices, revealing that our architecture can scale up to millions of records, in contrast to previous SOTA approaches, which are significantly slower.

Future Directions: Based on our findings and previous work, we envisage promising future research directions. To better guide semantic search and information retrieval while devising efficient and interpretable methods, we highlight the value of Neuro-Symbolic AI [36] and distributed learning [37]. Performances may skyrocket by infusing—jointly with raw texts and images—their unambiguous structured representations. The latter include semantic parsing graphs [38–40] or corpus-aware latent correlations for text [41–44] (e.g., with TF-IDF weighting [45] as node relevance score, link discovery via randomized perturbation [46,47]) and visual segmentation graphs for images [48]. Retrieval from external knowledge graphs and unstructured memories [49] can also act as a data augmentation strategy to cope with the lack of instances, usually addressed with transfer learning methods [50].

CRediT authorship contribution statement

Gianluca Moro: Conceptualization, Methodology, Investigation, Resources, Supervision, Writing - review & editing, Project administration. **Stefano Salvatori:** Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Writing - original draft. **Giacomo Frisoni:** Validation, Visualization, Writing - original draft.

Data availability

The datasets used in this article are publicly available

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research is partially supported by (i) the Complementary National Plan PNC-I.1 \quotes{Research initiatives for innovative technologies and pathways in the health and welfare sector} D.D. 931 of 06/06/2022, DARE---DigitAl lifelong pRevEntion initiative, code PNC0000002, CUP B53C22006450001, (ii) the PNRR---M4C2---Investment 1.3, Extended Partnership PE00000013, FAIR---Future Artificial Intelligence Research, Spoke 8 \quotes{Pervasive AI,} funded by the European Commission under the NextGeneration EU program.

Appendix A. Implementation Details

We implemented our model using *PyTorch* [51] starting from an existing implementation of the Transformer model available in the *HuggingFace's Transformers* library [52] to which we added the image processing layers and the Masked Patch Prediction head in order to compute the image reconstruction loss. For the metric learning phase, we used the *Pytorch Metric Learning* library [53].

The Transformer architecture consists of 12 layers with a 768 hidden size, 12 self-attention heads, and a 3072 intermediate size; we used `gelu` as the activation function and a 0.1 dropout factor. We started with pretrained weights loaded from the RoBERTa-base model. The Masked Patch Prediction head and the image pro-

cessing layers, which are not available in the traditional architecture, were randomly initialized from a 0 mean, 0.2 standard deviations normal distribution.

Appendix B. Training Setup

The models were trained using a single GeForce RTX 3090 GPU with 24 GB of available RAM. We used the same Adam optimizer in both training phases with parameters $\beta_1 = 0.95, \beta_2 = 0.999$, and weight decay $1e^{-4}$. Pretraining was run for 20 epochs with a batch size of 16 records and base learning rate of $2e^{-5}$ warmed up for the first 5000 steps and reduced during training using a cosine scheduling strategy. For the metric learning phase, the number of epochs was reduced to 10, and we set a constant learning rate of $2e^{-5}$. Each batch consisted of 16 products from which we extracted 16 descriptions and 16 images for a resulting batch size of 32 embeddings. For triplet loss we set $m = 1.0$ and employ a *semi-hard* mining strategy so that only triplets that satisfy the condition $d(x_a, x_p) < d(x_a, x_n) < d(x_a, x_p) + m$ were used. We set d to be the Euclidean distance. For Angular Loss we used $\alpha = 45^\circ$ and in Multi-similarity Loss we set $\alpha = 2, \beta = 40, \lambda = 0.5, \epsilon = 0.1$ and we set S_{ij} to be the cosine similarity.

Appendix C. Qualitative Evaluation

C.1. Source of Errors

To investigate why and when our model makes mistakes, we took the queries from the FashionGen test set for which it could not find the correct target among the top 10 results. Table C.6 shows the percentage of queries of a specific category incorrectly retrieved, both for Text-to-Image and Image-to-Text use cases. We note that there are some categories for which the model makes more "mistakes". We justify this behavior with the underrepresentation of these classes (e.g., BELTS & SUSPENDERS, FINE JEWELRY) in the training dataset, which limits the number of samples fed to the model.

Moreover, from qualitative and empirical analysis, we saw that the following situations are the other most common source of errors:

- Target or query images are photos of the product taken from the back (Fig. C.5). Since descriptions sometimes refer to frontal details, it is challenging for the model to recognize them when they are not visible in the image.
- Target or query images for Bags, Jewelry, and Glasses contain a person wearing them instead of having the object in the foreground (Fig. C.6).

Fig. C.5. Two examples of errors caused by the query image (top row) and the target image (bottom row) being photos taken from the back.

Table C.6
Percentage of queries of a specific category for which the target is not returned in the first 10 results (Out-of-@10). For each category, the value on top is for Text-to-Image retrieval, and the one on the bottom is for Image-to-Text.

Category	Out-of-@10	Category ↓	Out-of-@10 ↓
FLATS	2.99 7.81	SHIRTS	13.00 8.41
SCARVES	3.57 3.45	LOAFERS	14.29 12.50
BOOTS	4.22 2.47	SKIRTS	16.30 18.05
SANDALS	5.77 5.36	EYEWEAR	16.67 31.91
HEELS	6.45 3.13	PANTS	17.26 15.81
DRESSES	7.94 5.76	JEANS	17.28 15.85
TOPS	7.99 9.13	SUITS & BLAZERS	21.05 22.22
TOTE BAGS	8.57 12.9	SHOULDER BAGS	21.10 19.17
HATS	8.74 6.45	CLUTCHES & POUCHES	22.22 11.76
SNEAKERS	8.99 8.02	BAG ACCESSORIES	25.00 25.00
LACE UPS	10.00 11.11	JEWELRY	29.63 22.73
SHORTS	10.49 8.39	BACKPACKS	31.25 19.57
MESSENGER BAGS & SATCHELS	11.11 12.50	FINE JEWELRY	33.33 44.44
SWEATERS	11.27 11.14	POUCHES & DOCUMENT HOLDERS	50.00 —
JACKETS & COATS	12.83 15.30	BELTS & SUSPENDERS	55.56 —

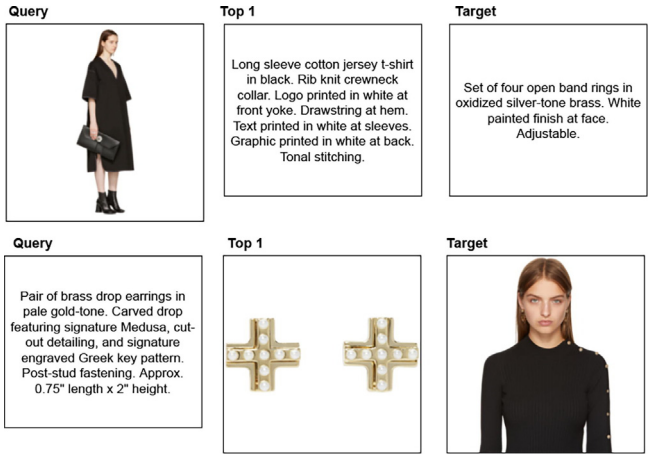


Fig. C.6. Two examples of errors caused by a query (top row) or target image (bottom row) that contains accessories worn by a person in the photo.

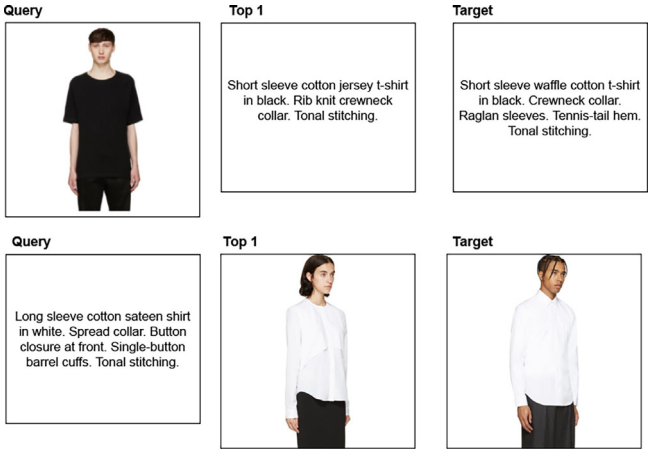


Fig. C.7. Two examples of failed queries due to the highly generic products. Despite being incorrect, the first result returned by the model is still similar to the target.

- The query text or image is too generic (Fig. C.7). Some akin clothes in the dataset, such as white t-shirts, have almost identical images and are often described similarly. Even if we seek a specific target, multiple results could be relevant when a query is too generic.

C.2. t-SNE Visualization

In Fig. C.8, we provide a low-dimensional representation of the latent space generated with our best model. We took the FashionGen validation set and run the t-SNE algorithm to project the 768-dimensional image embeddings into a 2D space. We used scikit-learn [54] implementation of t-SNE with the following parameters: `perplexity` = 65, `early_exaggeration` = 12.0, `n_iter` = 2000, `random_state` = 42, `learning_rate` = 200, `init`='pca'. We also embedded four random captions to show that they are placed near the images they describe and that the semantic relationship is preserved.

Appendix D. Ineffectiveness of FashionBERT in Managing Text-only and Image-only Queries

In the "Industry Application" section of the FashionBERT paper, the authors claim that their vision-language transformer can be used to encode text-only and image-only queries. Concretely, they say that their model can extract text and image embeddings by masking out all the image tokens or text tokens according to the type of input provided. However, the authors did not present any experiment to validate this idea, and we claim that it does not work in practice since the model is not trained to handle single texts or images in input. Furthermore, no other paper that uses a similar retrieval architecture (e.g., OSCAR, ViLBERT) suggests using this approach, not even KaleidoBERT. The efficiency issue of V + L transformers is, in fact, a problem that has also been studied and described in the work of Miech et al. [13].

To prove that the masking approach proposed in FashionBERT does not work, we tested it on the FashionGen dataset using our pretrained model and reported the results in Table D.7. As we can see, the Rank@*K* is only slightly above the accuracy of a random ranker; this shows that the latent representations extracted by masking out the tokens of one modality are not preserving the necessary information.

Appendix E. Web Application

In this section, we provide a brief description of the web application that we have released at https://disi-unibo-nlp.github.io/projects/fashion_retrieval/. The website can be used to perform text-to-image retrieval with both our model and FashionBERT[4] and compare their retrieval performance. The web application contains two pages: Search Description and Free Search.

Search Description Page: This page can be used to semantically search a product by its formal description (Fig. E.9). A web form allows selecting the following options.

- `Model`: the model to be used for retrieval (*our* or *FashionBERT*).
- `Candidate Set`: the dataset among which the product will be searched. Possible options are reported below. Please note that FashionBert requires pre-computed ResNet vectors as additional input, but authors released such data only for the validation images.
  - -Fixed Size Dataset: search a product among other 100 random ones from the same subcategory (this reproduces the test scenario used in this paper).
  - -Validation Dataset: search a product among the whole validation dataset (∼7500 products).
  - -Training Dataset: search a product among the whole training dataset (∼60000 products, only available if the selected model is *our*).
  - -Training + Validation: search a product using the full dataset (∼67500 products, only available if the selected model is *our*).
- `Product`: choose the product whose description will be used as a text query.
- `K-NN Strategy`: the nearest neighbor retrieval strategy (only available if model is *our*).
  - -In-Memory: use naïve nearest neighbor retrieval on embeddings loaded in memory.
  - -Postgres: use nearest neighbor retrieval on embeddings saved in a Postgres database.
  - -Index: use a multidimensional index for a faster search (note that you get the most significant speed-up when the largest dataset size is selected).

To test the difference in accuracy between our model and FashionBERT, try to search for the "Red Faded Plaid Shirt" product. First,

[4] To incorporate this model in the application, we used the code and pretrained weights available at https://github.com/alibaba/EasyTransfer/tree/master/scripts/fashion_bert

Fig. C.8. t-SNE visualization of the FashionGen validation set using image embeddings extracted with our model. 4 random captions are also embedded to show that semantic relationships are preserved. Best viewed in color at high resolution.

Table D.7
Results obtained performing retrieval on the FashionGen dataset using the pretrained model as suggested in the FashionBERT paper (i.e., masking out all image tokens or text tokens to encode text-only and image-only queries).

Model	TIR			ITR			*SumR*
	Rank@1	Rank@5	Rank@10	Rank@1	Rank@5	Rank@10	
Pretrain with masking	2.2%	6.6%	12.7%	3.6%	10.7%	20.0%	54.8

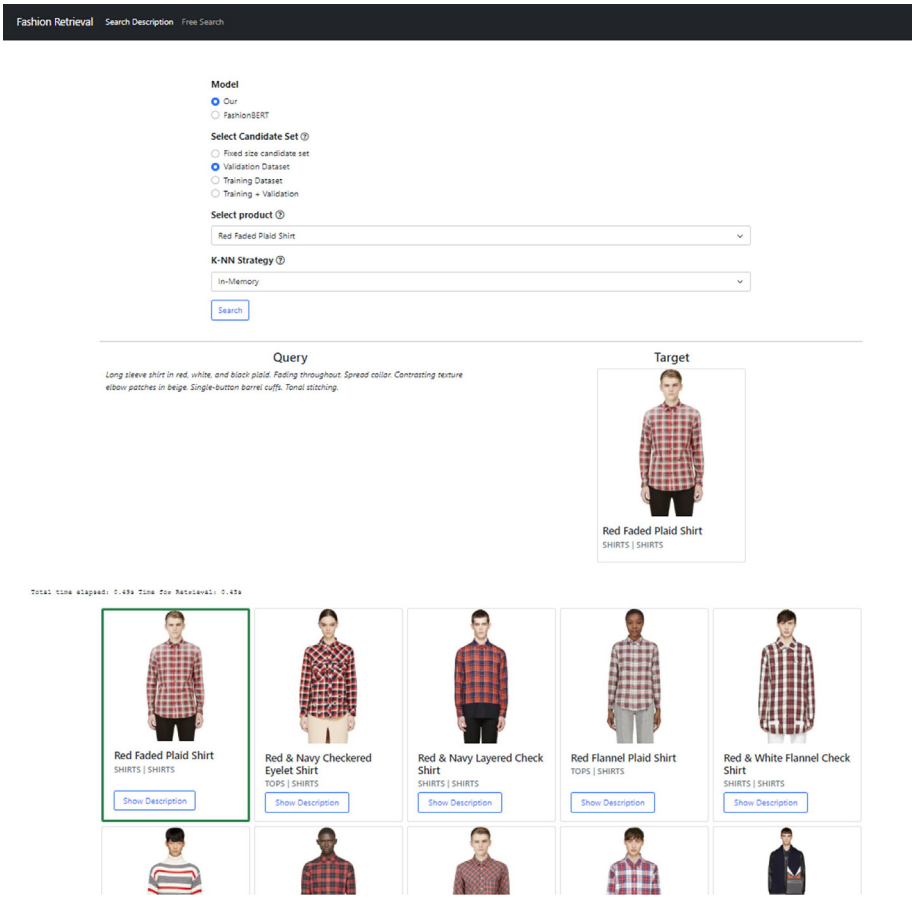


Fig. E.9. View of the Search Description Page on our Fashion Retrieval webapp.

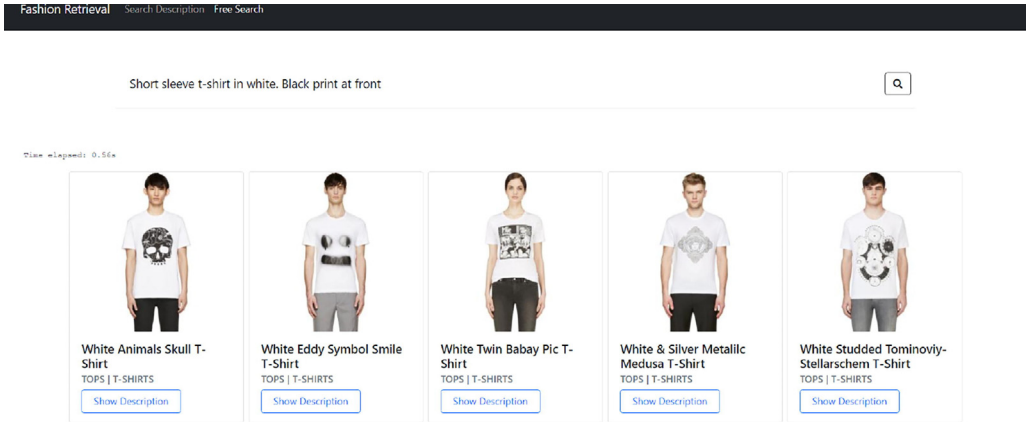


Fig. E.10. View of the Free Search Page on our Fashion Retrieval webapp.

select `Model = our`, `Candidate set = Validation dataset`, `Product = Red Faded Plaid Shirt` and `K-NN strategy = in memory`. If you click the "Search" button, you'll see that the correct target image, highlighted with a green border, is retrieved as the first result. If you switch from `Model = our` to `Model = fashion-bert`, you'll see that the target image is returned as the 18^{th} result. Furthermore, our model is more than 100x faster than FashionBERT ($\sim$0.5 s against $\sim$60 s) to find an image among almost 7000 products.

Free Search Page: The *Free Search* page (Fig. E.10) can be adopted to write a custom query and retrieve the most similar images using our model. The images are retrieved from the full FashionGen dataset (i.e., training + validation images).

We suggest trying the following queries to test the capabilities of our model:

- "Short sleeve t-shirt in white. Black print at front";
- "Long sleeve t-shirt in white. Red print at front";
- "Skinny-fit jeans in light blue".

References

[1] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, H. Wang, Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval, in: J. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J. Wen, Y. Liu (Eds.), Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25–30, 2020, ACM, 2020, pp. 2251–2260. doi:10.1145/3397271.3401430. URL: doi: 10.1145/3397271.3401430.

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S.V.N. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA, 2017, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[3] N. Rostamzadeh, S. Hosseini, T. Boquet, W. Stokowiec, Y. Zhang, C. Jauvin, C. Pal, Fashion-gen: The generative fashion dataset and challenge, CoRR abs/1806.08317. arXiv:1806.08317. URL: http://arxiv.org/abs/1806.08317.

[4] S. Zhu, S. Fidler, R. Urtasun, D. Lin, C.C. Loy, Be your own prada: Fashion synthesis with structural coherence, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, IEEE Computer Society, 2017, pp. 1689–1697. doi:10.1109/ICCV.2017.186. URL: doi: 10.1109/ICCV.2017.186.

[5] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, L. Shao, Kaleido-bert: Vision-language pre-training on fashion domain, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[6] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, A. Sacheti, Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data, CoRR abs/2001.07966. arXiv:2001.07966. URL: https://arxiv.org/abs/2001.07966

[7] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: pre-training of generic visual-linguistic representations, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SygXPaEYvH.

[8] G. Li, N. Duan, Y. Fang, M. Gong, D. Jiang, Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training, in: AAAI, AAAI Press, 2020, pp. 11336–11344.

[9] X. Li, X. Yin, C. Li, X. Hu, P. Zhang, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, Oscar: Object-semantics aligned pre-training for vision-language tasks, ECCV (2020).

[10] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, 2019, pp. 13–23. URL: https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

[11] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, IEEE Computer Society, 2016, pp. 770–778. doi:10.1109/CVPR.2016.90. URL: doi: 10.1109/CVPR.2016.90.

[12] G. Ji, M. Zhuge, D. Gao, D. Fan, C. Sakaridis, L.V. Gool, Masked vision-language transformer in fashion, CoRR abs/2210.15110.

[13] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, A. Zisserman, Thinking fast and slow: Efficient text-to-visual retrieval with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 9826–9836.

[14] S. Mirchandani, L. Yu, M. Wang, A. Sinha, W. Jiang, T. Xiang, N. Zhang, Fad-vlp: Fashion vision-and-language pre-training towards unified retrieval and captioning, CoRR abs/2210.15028.

[15] X. Han, L. Yu, X. Zhu, L. Zhang, Y. Song, T. Xiang, Fashionvil: Fashion-focused vision-and-language representation learning, in: ECCV (35), Vol. 13695 of Lecture Notes in Computer Science Springer, 2022, pp. 634–651.

[16] L. Yu, J. Chen, A. Sinha, M. Wang, Y. Chen, T.L. Berg, N. Zhang, Commercemm: Large-scale commerce multimodal representation learning with omni retrieval, in: KDD, ACM, 2022, pp. 4433–4442.

[17] H. Jégou, M. Douze, C. Schmid, Product quantization for nearest neighbor search, IEEE Trans. Pattern Anal. Mach. Intell. 33 (1) (2011) 117–128, https://doi.org/10.1109/TPAMI.2010.57.

[18] G. Moro, L. Valgimigli, Efficient self-supervised metric information retrieval: A bibliography based method applied to COVID literature, Sensors 21 (19). doi:10.3390/s21196430. URL: doi: 10.3390/s21196430.

[19] G. Moro, L. Ragazzi, Semantic Self-Segmentation for Abstractive Summarization of Long Legal Documents in Low-Resource Regimes, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Virtual Event, February 22 - March 1, 2022, AAAI Press, 2022, pp. 1–9.

[20] A. Frome, G.S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, T. Mikolov, Devise: A deep visual-semantic embedding model, in: C.J.C. Burges, L. Bottou, Z. Ghahramani, K.Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States, 2013, pp. 2121–2129. URL: https://proceedings.neurips.cc/paper/2013/hash/7cce53cf90577442771720a370c3c723-Abstract.html

[21] E. Hoffer, N. Ailon, Deep metric learning using triplet network, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Workshop Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6622.

[22] F. Faghri, D.J. Fleet, J.R. Kiros, S. Fidler, VSE++: improving visual-semantic embeddings with hard negatives, in: British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3–6, 2018, BMVA Press, 2018, p. 12. URL: http://bmvc2018.org/contents/papers/0344.pdf.

[23] K. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: ECCV (4), Vol. 11208 of Lecture Notes in Computer Science Springer, 2018, pp. 212–228.

[24] Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, X. Fan, Position focused attention network for image-text matching, in: S. Kraus (Ed.), Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10–16, 2019, ijcai.org, 2019, pp. 3792–3798. doi:10.24963/ijcai.2019/526. URL: https://doi.org/10.24963/ijcai.2019/526.

[25] S. Ren, K. He, R.B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, CoRR abs/1506.01497. arXiv:1506.01497. URL: http://arxiv.org/abs/1506.01497

[26] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event, Vol. 139 of Proceedings of Machine Learning Research, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[27] P.J. Chia, G. Attanasio, F. Bianchi, S. Terragni, A.R. Magalhães, D. Goncalves, C. Greco, J. Tagliabue, Fashionclip: Connecting language and images for product representations, CoRR abs/2204.03972. arXiv:2204.03972, doi:10.48550/arXiv.2204.03972. URL: https://doi.org/10.48550/arXiv.2204.03972.

[28] W. Shin, J. Park, T. Woo, Y. Cho, K. Oh, H. Song, e-clip: Large-scale vision-language representation learning in e-commerce, in: CIKM, ACM, 2022, pp. 3484–3494.

[29] H. Liu, S. Xu, J. Fu, Y. Liu, N. Xie, C. Wang, B. Wang, Y. Sun, CMA-CLIP: cross-modality attention CLIP for image-text classification, CoRR abs/2112.03562.

[30] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692. arXiv:1907.11692. URL: http://arxiv.org/abs/1907.11692

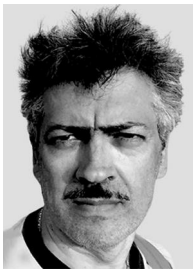
[31] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021, OpenReview.net, 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[32] J. Wang, F. Zhou, S. Wen, X. Liu, Y. Lin, Deep metric learning with angular loss, in: IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22–29, 2017, IEEE Computer Society, 2017, pp. 2612–2620. doi:10.1109/ICCV.2017.283. URL: doi: 10.1109/ICCV.2017.283.

[33] X. Wang, X. Han, W. Huang, D. Dong, M.R. Scott, Multi-similarity loss with general pair weighting for deep metric learning, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, Computer Vision Foundation/ IEEE, 2019, pp. 5022–5030. doi:10.1109/CVPR.2019.00516. URL: http://openaccess.thecvf.com/content_CVPR_2019/html/Wang_Multi-Similarity_Loss_With_General_Pair_Weighting_for_Deep_Metric_Learning_CVPR_2019_paper.html.

[34] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlós, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, D. Belanger, L. Colwell, A. Weller,

Rethinking attention with performers, in: International Conference on Learning Representations, ICLR 2021, 2021.
[35] G. Moro, L. Ragazzi, L. Valgimigli, D. Freddi, Discriminative marginalized probabilistic neural method for multi-document summarization of medical literature, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 180–189. doi:10.18653/v1/2022.acl-long.15. URL: https://aclanthology.org/2022.acl-long.15.
[36] G. Frisoni, G. Moro, A. Carbonaro, Towards Rare Disease Knowledge Graph Learning from Social Posts of Patients, in: RiiForum, Springer, 2020, pp. 577–589, https://doi.org/10.1007/978-3-030-62066-0_44, https://www.scopus.com/inward/record.uri?eid=2-s2.0-85102640128&doi=10.1007%2f978-3-030-62066-0_44&partnerID=40&md5=7b08bda5b0f9de00d4e5acdaccfe7707s.
[37] W. Cerroni, G. Moro, T. Pirini, M. Ramilli, Peer-to-peer data mining classifiers for decentralized detection of network attacks, in: H. Wang, R. Zhang (Eds.), Twenty-Fourth Australasian Database Conference, ADC 2013, Adelaide, Australia, February 2013, Vol. 137 of CRPIT, Australian Computer Society, 2013, pp. 101–108. URL: http://crpit.scem.westernsydney.edu.au/abstracts/CRPITV137Cerroni.html.
[38] G. Frisoni, G. Moro, A. Carbonaro, Unsupervised Descriptive Text Mining for Knowledge Graph Learning, in: IC3K 2020 - Proc. 12th Int. Joint Conf. Knowl. Discovery, Knowl. Eng. and Knowl. Manage., Vol. 1, SciTePress, 2020, pp. 316–324. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85107113340&partnerID=40&md5=7a4cc3ae8a6894d1a3fff499bb4bf717.
[39] G. Frisoni, G. Moro, A. Carbonaro, A survey on event extraction for natural language understanding: Riding the biomedical literature wave, IEEE Access 9 (2021) 160721–160757, https://doi.org/10.1109/ACCESS.2021.3130956.
[40] G. Frisoni, G. Moro, G. Carlassare, A. Carbonaro, Unsupervised event graph representation and similarity learning on biomedical literature, Sensors 22 (1) (2022) 3, https://doi.org/10.3390/s22010003.
[41] G. Frisoni, G. Moro, A. Carbonaro, Learning Interpretable and Statistically Significant Knowledge from Unlabeled Corpora of Social Text Messages: A Novel Methodology of Descriptive Text Mining, in: DATA 2020 - Proc. 9th Int. Conf. Data Science, Technol. and Appl., SciTePress, 2020, pp. 121–134. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85092009636&partnerID=40&md5=27541a3b46d782bb7984eed8ba7fa8a3.
[42] G. Frisoni, G. Moro, Phenomena Explanation from Text: Unsupervised Learning of Interpretable and Statistically Significant Knowledge, in: DATA (Revised Selected Papers), Vol. 1446, Springer, 2020, pp. 293–318. doi:10.1007/978-3-030-83014-4_14. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85113292013&doi=10.1007%2f978-3-030-83014-4_14&partnerID=40&md5=33fa92fd1f11dff84de31aac3729917a.
[43] G. Domeniconi, K. Semertzidis, V. López, E.M. Daly, S. Kotoulas, G. Moro, A novel method for unsupervised and supervised conversational message thread detection, in: DATA 2016 - Proc. 5th Int. Conf. Data Science, Technol. and Appl., Lisbon, Portugal, 24–26 July, 2016, SciTePress, 2016, pp. 43–54. doi:10.5220/0006001100430054. URL: doi: 10.5220/0006001100430054.
[44] G. Domeniconi, G. Moro, A. Pagliarani, K. Pasini, R. Pasolini, Job Recommendation from Semantic Similarity of LinkedIn Users' Skills, in: ICPRAM 2016, SciTePress, 2016, pp. 270–277. doi:10.5220/0005702302700277. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84970039381&doi=10.5220%2f0005702302700277&partnerID=40&md5=eca4633aae1e9418df034aaa5f3a6020.
[45] G. Domeniconi, G. Moro, R. Pasolini, C. Sartori, A Comparison of Term Weighting Schemes for Text Classification and Sentiment Analysis with a Supervised Variant of tf.idf, in: DATA (Revised Selected Papers), Vol. 584, Springer, 2015, pp. 39–58. doi:10.1007/978-3-319-30162-4_4. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84961127206&doi=10.1007%2f978-3-319-30162-4_4&partnerID=40&md5=81e9a8dc2045e1186bf840b7e43e3118.
[46] G. Domeniconi, M. Masseroli, G. Moro, P. Pinoli, Discovering new gene functionalities from random perturbations of known gene ontological annotations, INSTICC Press (2014) 107–116, https://doi.org/10.5220/0005087801070116, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84909957332&doi=10.5220%2f0005087801070116&partnerID=40&md5=d46ef212e92f6a5b1c3d3769ca8a0564.
[47] G. Domeniconi, M. Masseroli, G. Moro, P. Pinoli, Cross-organism learning method to discover new gene functionalities, Comput. Methods Programs Biomed. 126 (2016) 20–34, https://doi.org/10.1016/j.cmpb.2015.12.002.
[48] S. Sadegharmaki, M.A. Kastner, S. Satoh, Fashiongraph: Understanding fashion data using scene graph generation, in: 25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event/ Milan, Italy, January 10–15, 2021, IEEE, 2020, pp. 7923–7929. doi:10.1109/ICPR48806.2021.9412662. URL: doi: 10.1109/ICPR48806.2021.9412662.
[49] G. Moro, A. Pagliarani, R. Pasolini, C. Sartori, Cross-domain & In-domain Sentiment Analysis with Memory-based Deep Neural Networks, in: IC3K 2018, Vol. 1, SciTePress, 2018, pp. 127–138. doi:10.5220/0007239101270138. URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85059000370&doi=10.5220%2f0007239101270138&partnerID=40&md5=257a04cbdf98a4d75275d39563b0aa17.
[50] G. Domeniconi, G. Moro, R. Pasolini, C. Sartori, Iterative Refining of Category Profiles for Nearest Centroid Cross-Domain Text Classification, in: IC3K 2014, Rome, Italy, October 21–24, 2014, Revised Selected Papers, Vol. 553, Springer, 2014, pp. 50–67. doi:10.1007/978-3-319-25840-9_4. URL: doi: 10.1007/978-3-319-25840-9_4.
[51] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, Pytorch: An imperative style, high-performance deep learning library, in: H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8–14, 2019, Vancouver, BC, Canada, 2019, pp. 8024–8035. URL: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.
[52] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 38–45.
[53] K. Musgrave, S. Belongie, S.-N. Lim, Pytorch metric learning (2020). arXiv:2008.09164.
[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. VanderPlas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830, URL http://dl.acm.org/citation.cfm?id=2078195.



Gianluca Moro received the M.S. degree in computer science from the University of Bologna, Italy, in 1994 and the Ph.D. degree in computer engineering at the Department of Electronics, Computer Science and Systems, Italy, in 1999. He is Senior Research Associate at the Department of Computer Science and Engineering of the University of Bologna and is professor of text mining, data mining and big data analytics. He is head of the research unit in text mining and natural language processing at Cesena campus. He co-organized several editions of workshops at VLDB and AAMAS, edited five international books, and published more than ninety papers, some awarded as best papers in data science and information retrieval conferences. He has leaded national and international projects on data mining and machine learning research topics and collaborates with IBM Thomas J. Watson Research Center, US. He has the qualification of associate professor approved by the Italian Minister of University and Research.

Stefano Salvatori received the B.S. degree and M.S. degree with honors in computer science and engineering from the University of Bologna, Italy, in 2018 and 2021. Currently, he is a research assistant for the Department of Computer Science and Engineering (DISI) in Cesena. His research interests include natural language processing, information retrieval, and vision-language models and their applications.

Giacomo Frisoni received the B.S. and M.S. degrees in computer science and engineering from the University of Bologna, Italy, in 2017 and 2020, respectively-both with honors. He is currently a second year Ph.D. student at the Department of Computer Science and Engineering, University of Bologna. His research interests include natural language understanding, neuro-symbolic AI, and graph neural networks. He presented several original papers to international journals and peer-reviewed conferences–including top-tier venues like EMNLP and COLING, winning two Best Paper Awards. In 2020, he was among the worldwide selected program attendees at the Cornell, Maryland, Max Planck Pre-doctoral Research School. In the same year, he received the con. Science Award for having written one of the ten Italian best scientific research works during the master's thesis. In June 2022, he was selected as a member of the first Hugging Face Student Ambassador program.