4-2021

Sentiment-oriented metric learning for text-to-image retrieval

Quoc Tuan TRUONG
Singapore Management University, qttruong@smu.edu.sg

Hady W. LAUW
Singapore Management University, hadywlauw@smu.edu.sg

Citation

TRUONG, Quoc Tuan and LAUW, Hady W.. Sentiment-oriented metric learning for text-to-image retrieval. (2021). *Advances in Information Retrieval: 43rd European Conference on IR Research ECIR 2021, Virtual, March 28 - April 1: Proceedings*. 12656, 634-649.
Available at: https://ink.library.smu.edu.sg/sis_research/5951

Sentiment-Oriented Metric Learning for Text-to-Image Retrieval

Quoc-Tuan Truong and Hady W. Lauw$^{[0000-0002-8245-8677]}$

Singapore Management University, Singapore
{qttruong.2017,hadywlauw}@smu.edu.sg

Abstract. In this era of multimedia Web, text-to-image retrieval is a critical function of search engines and visually-oriented online platforms. Traditionally, the task primarily deals with matching a text query with the most relevant images available in the corpus. To an increasing extent, the Web also features visual expressions of preferences, imbuing images with sentiments that express those preferences. Cases in point include photos in online reviews as well as social media. In this work, we study the effects of sentiment information on text-to-image retrieval. Particularly, we present two approaches for incorporating sentiment orientation into metric learning for cross-modal retrieval. Each model emphasizes a hypothesis on how positive and negative sentiment vectors may be aligned in the metric space that also includes text and visual vectors. Comprehensive experiments and analyses on *Visual Sentiment Ontology (VSO)* and *Yelp.com* online reviews datasets show that our models significantly boost the retrieval performance as compared to various sentiment-insensitive baselines.

Keywords: Text-to-Image Retrieval · Cross-Modal Retrieval · Metric Learning · Sentiment Orientation.

1 Introduction

The Web is awash in visual imagery. Millions of images are added daily to the billions already existing in various image-oriented platforms such as Instagram, Pinterest, Flickr, etc. In addition, product reviews in virtually any category, be it of restaurants on Yelp or consumer electronics on Amazon, frequently feature photos accompanying (complementing and even enhancing) the textual content of the reviews. In the face of such abundance and diversity, finding images relevant to one's purpose remains a pertinent challenge. While images are now a cornerstone modality on the Web, the manner in which most users express their intent is still predominantly textual. In this paper, we focus on text-to-image retrieval, i.e., retrieving images from a textual query. This is distinct from image retrieval, i.e., retrieving images from an image query [10], which is an active research topic in its own right.

The presumption by many previous works on cross-modal retrieval (involving multiple modalities, such as text and image) [8, 41] is that queries, and by extension the images the queries are aimed at, are generally of an objective nature.

For instance, a user may be looking for pictures of a cat, a car, a specific person, etc. In reality, images are not universally devoid of sentiment. To the contrary, recent literature on visual sentiment analysis [47, 36–38, 31] attests to the manifestation of sentiments within some images. Within reviews for a restaurant or a hotel for example, someone may post an image of "restroom" in the positive sense (perhaps an especially clean or well-appointed specimen) or in the negative sense (such as the case where hygiene is less than desired). Conceivably, an "objective" query may turn out images of varied sentiments, due to its lack of specificity of which sentiment is fit and proper for the occasion at hand.

Problem. For a more holistic and expressive capacity for retrieving relevant images, we posit that in some scenarios the query intent may indeed have a sentiment dimension. For simplicity of discourse, we assume binary sentiment classes of *positive* and *negative* respectively. In other words, a query is now a tuple of *(textual keywords, sentiment class)*, and we seek to return a ranked list of images (from a corpus), which are relevant to the specified keywords *and* sentiment. It is worth noting that the corpus of interest consists of mere images, unadorned explicitly with text nor sentiment.

There are several challenges to this problem. One challenge inherent to cross-modality learning is how to learn associations among different modalities with distinct feature spaces, in this case text and images. Another challenge pertinent to retrieval is how to model relevance between varied modalities. Over and above these that plague cross-modal retrieval, we also have the peculiar challenge of modeling the third modality of interest, namely sentiment.

Approach. To deal with these challenges, we propose a framework called *Sentiment-Oriented Metric Learning* or *SML*. To overcome the variety in modalities, we learn modality-specific feature mappings that respectively map text and image inputs onto a common space. Presuming training data in the form of *text-sentiment-image* triples, we preserve relevant associations in these triples through proximity constraints relating texts, sentiments, as well as images in the resulting common feature space. Of particular interest are the manners in which we model sentiments as directional vectors in the common metric space, giving rise to two variants, SML_{OPPO} and SML_{FLEX}, based on different assumptions in bringing sentiment-oriented queries closer to the relevant images.

Contributions. In this work, we make several contributions. First, to our best awareness, this is the first work to study the effect of sentiment information for better understanding of text-to-image retrieval. Second, to characterize the effect of sentiments, we develop two models, namely SML_{OPPO} and SML_{FLEX}, that learn metric spaces in which the sentiments are represented by directional vectors. Third, we conduct comprehensive experiments comparing the proposed models with other cross-modal retrieval approaches. Experiments on real-life datasets, which include *Visual Sentiment Ontology* (Flickr images) and online review images from *Yelp.com*, show that our models significantly outperform the sentiment-insensitive baselines, underlining the import of sentiment on text-to-image retrieval.

2 Related Work

In this section, we review the related work along the two broad lines of metric learning as well as multi-view learning.

Metric Space Learning. The notion of distance is fundamental to many machine learning algorithms. Metric representation learning [22] deals with learning representations of objects so as to reflect the relationships among those objects in terms of distances in the metric space, i.e., putting relevant objects in proximity while distancing irrelevant ones. It finds applications in various contexts, such as image classification [35, 26], image retrieval [40, 21], text retrieval [46] and collaborative filtering [15], whereby in each case context-specific constraints may apply.

In the context of cross-modal retrieval, the constraints may include minimizing distances between positive pairs while maximizing distances between negative pairs [25, 24]. Additional considerations may include preserving geometric structures such as global consistency and local smoothness [48] or making the feature learning modality-specific [42, 49]. Orthogonally, we investigate metric learning for sentiment-oriented text-to-image retrieval, whereby the sentiment-orientation is particularly novel. We further propose a framework incorporating recent developments in deep representation learning, with new objectives to factor sentiment into the learned metric space (in addition to text and image modalities).

Multi-View Learning. An object may have multiple "views", i.e., observations in distinct feature spaces. In cross-modal retrieval, we have text and images. Multi-view learning finds object representations across several views, which would preserve the associations among different views of an object as well as among objects within a view.

A classical technique for feature learning across spaces is Canonical Correlation Analysis (CCA) [14, 2]. The crux is to find linear projections of two vectors (one for each view), so as to maximize their correlations. To incorporate nonlinearity, one approach is based on kernel methods [1, 27, 5, 11]. A more recent approach is to use deep neural networks [23, 16, 3], of which DCCA [3] is the most recent work presenting a complete learning framework. In experiments, we compare to both CCA and DCCA.

Aside from correlation analysis, neural networks are also used for multi-view learning in different ways. Within the autoencoder framework, the objective is usually to find a feature representation in a common space that could reconstruct the inputs in the respective feature spaces [29, 8, 43]. In turn, [30, 41] employ adversarial learning framework. As a recent competitive method for cross-modal retrieval based on adversarial learning, ACMR [41] is included as a baseline.

Note that ours has a different problem setting from those [32, 17, 41] that learn discriminative common representations by exploiting labels to distinguish between semantic categories. For one, sentiment can be seen as an independent modality, rather than labels during learning. For another, sentiment itself is a part of the query. Also incidentally related are approaches based on cross-modal hashing [33, 34, 45, 7, 50] that focus primarily on retrieval efficiency, while tolerating some loss in accuracy due to potential loss of information.

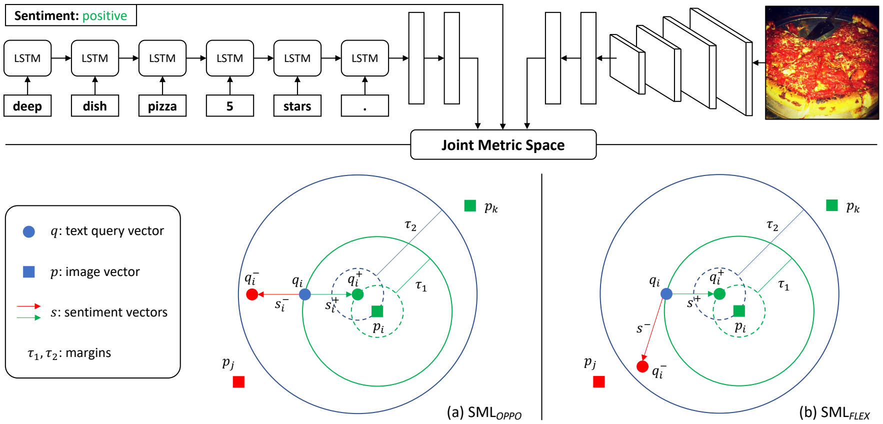

Fig. 1: Illustration of the SML framework. Image and text are embedded into the metric space using deep neural networks. For **SML**$_{OPPO}$ (a), sentiment vectors are in opposite directions, while sentiment vectors in **SML**$_{FLEX}$ (b) are unconstrained. Given that the query is *positive*, the sentiment margin constraint, $d(\mathbf{q}_i^+, \mathbf{p}_i) < d(\mathbf{q}_i, \mathbf{p}_i) - \tau_1$, is demonstrated in green color (*negative* is in red color). In turn, the distance margin constraint between correct and incorrect query-photo pairs, $d(\mathbf{q}_i^+, \mathbf{p}_i) < d(\mathbf{q}_i^+, \mathbf{p}_j) - \tau_2$, is demonstrated in blue color.

3 Sentiment-Oriented Metric Learning (SML)

An input data collection $\mathcal{T} = \{(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i)\}_{i=1}^N$ contains N instances of *text-sentiment-image* triples. Here, $\mathbf{z}_i$ is binary $\{$*positive*, *negative*$\}$. Our objective is to infuse the text with sentiment in order to form a sentiment-sensitive query $(\mathbf{x}_i, \mathbf{z}_i)$ that would better match the desired image $\mathbf{y}_i$ than $\mathbf{x}_i$ could on its own.

In essence, we propose SML framework which seeks to find two functions f and g transforming queries and images, respectively, into a metric space in which their similarities can be measured. Specifically, $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\psi}}$, parameterized by $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, independently map $(\mathbf{x}_i, \mathbf{z}_i)$ and $\mathbf{y}_i$ to a D-dimensional Euclidean space $\mathbb{R}^D$, in which the distance between query $(\mathbf{x}_i, \mathbf{z}_i)$ and image $\mathbf{y}_i$ is measured as:

$$d_{\boldsymbol{\theta},\boldsymbol{\psi}}((\mathbf{x}_i, \mathbf{z}_i), \mathbf{y}_i) = \|f_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{z}_i) - g_{\boldsymbol{\psi}}(\mathbf{y}_i)\|_2 \tag{1}$$

In this framework, we posit that sentiments are high-level abstraction concepts which should be represented as independent vectors in the metric space. We model a sentiment-infused query in additive form $f_{\boldsymbol{\theta}}(\mathbf{x}_i, \mathbf{z}_i) = \mathbf{q}_i + \mathbf{s}_i$, where $\mathbf{q}_i$ and $\mathbf{s}_i$ are vectors in the metric space representing text and sentiment respectively. In turn, $g_{\boldsymbol{\psi}}(\mathbf{y}_i) = \mathbf{p}_i$ is a vector representing the image in the same metric space. The specific instantiation of $\mathbf{s}_i$ manifests slightly differently in two models, SML$_{OPPO}$ and SML$_{FLEX}$, which will be discussed subsequently. The learning output consists of transformations for $\mathbf{q}_i$ and $\mathbf{p}_i$, as well as the sentiment vectors $\{\mathbf{s}_i\}$ that allow us to measure distances for new queries and images.

3.1 Opposing Sentiment Vectors (SML$_{OPPO}$)

In the first model, referred to as SML$_{OPPO}$, we propose learning opposing sentiment vectors in a metric space. In other words, the two sentiment vectors (*positive* and *negative*) are in opposite directions and having the same magnitude. Thus, we only need to learn a single vector $\mathbf{s}$. It follows that the positive vector is $+\mathbf{s}$ and the negative vector is $-\mathbf{s}$. For each query tuple $(\mathbf{x}_i, \mathbf{z}_i)$, the sentiment vector $\mathbf{s}_i$ is in the form of:

$$\mathbf{s}_i = \alpha_i * \Gamma(\mathbf{z}_i) * \mathbf{s} \tag{2}$$

$$\alpha_i = \ln(1 + \exp(\mathbf{W}_\alpha^T \mathbf{q}_i)) \tag{3}$$

$$\Gamma(\mathbf{z}_i) = \begin{cases} +1 & \text{if } \mathbf{z}_i = \textit{positive} \\ -1 & \text{if } \mathbf{z}_i = \textit{negative} \end{cases} \tag{4}$$

where $\mathbf{s}$ is the sentiment basis vector shared across queries, $\Gamma(*)$ is a sign function, α_i is query-specific scale factor controlling the magnitude of the sentiment vector $\mathbf{s}_i$. Hypothetically, α_i is a function of $\mathbf{q}_i$ as different semantic concepts in different text queries require different intensity for the sentiment to be expressed. The choice of *softplus* [9] function for α_i is because of its smoothness and to ensure the value domain $\alpha_i \in (0, +\infty)$ for vector magnitude.

Our model learning can be specified as a *constrained* optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{W}_\alpha, \mathbf{s}} \quad & \lambda(r(\boldsymbol{\theta}) + r(\boldsymbol{\psi})) + \sum_{i=1}^{N} d(\mathbf{q}_i + \mathbf{s}_i, \mathbf{p}_i) \\ \text{s.t.} \quad & d(\mathbf{q}_i + \mathbf{s}_i, \mathbf{p}_i) < d(\mathbf{q}_i, \mathbf{p}_i) - \tau_1 \\ & d(\mathbf{q}_i + \mathbf{s}_i, \mathbf{p}_i) < d(\mathbf{q}_i + \mathbf{s}_i, \mathbf{p}_j) - \tau_2, \forall j \neq i \end{aligned} \tag{5}$$

where $r(*)$ is regularizer on the model parameters $\{\boldsymbol{\theta}, \boldsymbol{\psi}\}$, $d(*)$ is the loss due to Euclidean distance, and λ is the trade-off between regularizer and loss. The first constraint is margined relative distance between sentiment-oriented-query and neutral-query towards the correct image. The second constraint is margined relative distance between correct and incorrect query-photo pairs. The relationships amongst vectors and constraints are demonstrated in Fig. 1a.

We transform this constrained optimization into a regularized empirical risk minimization problem. The constraints are enforced using the standard hinge loss $[\delta]_+ = \max(0, \delta)$. We then derive an unconstrained loss function with $l2$-regularization as follows:

$$\begin{aligned} \mathcal{L} = \lambda\Big(\|\mathbf{W}_\alpha\|_F^2 + \sum_{l=1}^{L_f}(\|\mathbf{W}_f^l\|_F^2 + \|\mathbf{b}_f^l\|_2^2) + \sum_{l=1}^{L_g}(\|\mathbf{W}_g^l\|_F^2 + \|\mathbf{b}_g^l\|_2^2)\Big) \\ + \sum_{i=1}^{N}\Big[\|(\mathbf{q}_i + \mathbf{s}_i) - \mathbf{p}_i\|_2^2 + \max\big(0, \tau + \|(\mathbf{q}_i + \mathbf{s}_i) - \mathbf{p}_i\|_2^2 - \|\mathbf{q}_i - \mathbf{p}_i\|_2^2\big) \\ + \sum_{j=1}^{N} \mathbb{1}(i \neq j) \max\big(0, 1 + \|(\mathbf{q}_i + \mathbf{s}_i) - \mathbf{p}_i\|_2^2 - \|\mathbf{q}_i - \mathbf{p}_j\|_2^2\big)\Big] \end{aligned} \tag{6}$$

where L_f and L_g are the numbers of layers of the two neural networks characterizing $f_{\boldsymbol{\theta}}$ and $g_{\boldsymbol{\psi}}$, respectively.

Parameters of the model can be optimized via minimizing the loss function using stochastic gradient descent. In practice, we optimize the model using mini-batch to speed up the learning process. For each mini-batch of triples $\mathcal{B} = \{(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i)\}_{i=1}^{|\mathcal{B}|}$ sampled from the collection $\mathcal{T}$, each query will be paired with other images within the mini-batch to form negative pairs instead of considering all possible negative combinations from the whole collection $\mathcal{T}$. This stochastic process drastically reduces convergence time, and in expectation achieves our global objective (Eq. 6). Algorithm 1 describes the optimization procedure with the mini-batch gradient descent.

3.2 Flexible Sentiment Vectors (SML$_{FLEX}$)

In some ways, the previous assumption by SML$_{OPPO}$ could be quite restrictive, as the opposing directions of the sentiment vectors are enforced on every single dimension of the learned metric space.

To allow for greater flexibility, we arrive at another variant, which we refer to as SML$_{FLEX}$, by allowing the positive sentiment vector and negative sentiment vector to take their own independent directions. That way, they can be opposing in some dimensions, but not necessarily across all D dimensions. Thus, it provides another degree of freedom for the model to allocate coordinates judiciously between the objective of capturing sentimental concepts as well as that of capturing textual-visual semantic concepts. SML$_{FLEX}$ decouples and learns two global sentiment vectors $\mathbf{s}^+$ and $\mathbf{s}^-$ separately. Fig. 1b illustrates the learned metric space of SML$_{FLEX}$. The constrained optimization is as follows:

$$\begin{aligned} \min_{\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{s}^+, \mathbf{s}^-} \quad & \lambda(r(\boldsymbol{\theta}) + r(\boldsymbol{\psi})) + \sum_{i=1}^{N} d(\mathbf{q}_i + \mathbf{s}^{\Gamma(\mathbf{z}_i)}, \mathbf{p}_i) \\ \text{s.t.} \quad & d(\mathbf{q}_i + \mathbf{s}^{\Gamma(\mathbf{z}_i)}, \mathbf{p}_i) < d(\mathbf{q}_i, \mathbf{p}_i) - \tau_1 \\ & d(\mathbf{q}_i + \mathbf{s}^{\Gamma(\mathbf{z}_i)}, \mathbf{p}_i) < d(\mathbf{q}_i + \mathbf{s}^{\Gamma(\mathbf{z}_i)}, \mathbf{p}_j) - \tau_2, \forall j \neq i \end{aligned} \tag{7}$$

Similarly to SML$_{OPPO}$, we can derive an unconstrained loss function and proceed minimization with the stochastic gradient descent algorithm.

3.3 Implementation Details

In this work, we use two neural networks, recurrent and convolution, to learn text and image transformations. The former uses LSTM cell [13], which had been shown to be effective in learning textual representation in many machine learning tasks. Word embeddings to the LSTM are initialized from pre-trained Word2vec [28] of 300 dimensions. For the latter, we employ ResNet-50 [12] architecture, which has also been used extensively for obtaining image representation of numerous vision-related tasks. The output representations from LSTM and ResNet-50 are both projected into the metric space using two-layer perceptrons (each layer is followed by the *hyperbolic tangent* activation function). The implementation of SML is made available at *https://code.preferred.ai/sml/*.

Algorithm 1 Parameter learning with mini-batch gradient descent

```
Input: $\mathcal{T} = \{(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i)\}_{i=1}^{N}$, learning rate $\eta$
Output: Learned parameters $\{\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{s}\}$

initialization
    $\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{s} \leftarrow$ randomly initialized
while not converged do
    $\mathcal{T}_{batch} = \{\mathcal{B}_b\}_{b=1}^{num\_batch} \leftarrow$ uniformly sampled from $\mathcal{T}$
    for all $\mathcal{B}_b \in \mathcal{T}_{batch}$ do
        $g\boldsymbol{\theta} = 0;\ g\boldsymbol{\psi} = 0;\ g\mathbf{s} = 0;$
        for all $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i) \in \mathcal{B}_b$ do
            for all $(\mathbf{x}_j, \mathbf{z}_j, \mathbf{y}_j) \in \mathcal{B}_b$ where $(j \neq i)$ do
                $g\boldsymbol{\theta} = g\boldsymbol{\theta} + \frac{\partial}{\partial \boldsymbol{\theta}}\mathcal{L}(\mathbf{x}_i, \mathbf{s}_i, \mathbf{y}_i, \mathbf{y}_j);$
                $g\boldsymbol{\psi} = g\boldsymbol{\psi} + \frac{\partial}{\partial \boldsymbol{\psi}}\mathcal{L}(\mathbf{x}_i, \mathbf{s}_i, \mathbf{y}_i, \mathbf{y}_j);$
                $g\mathbf{s} = g\mathbf{s} + \frac{\partial}{\partial \mathbf{s}}\mathcal{L}(\mathbf{x}_i, \mathbf{s}_i, \mathbf{y}_i, \mathbf{y}_j);$
            end for
        end for
        $\boldsymbol{\theta} = \boldsymbol{\theta} - \eta \cdot \frac{g\boldsymbol{\theta}}{|\mathcal{B}_b|};\ \boldsymbol{\psi} = \boldsymbol{\psi} - \eta \cdot \frac{g\boldsymbol{\psi}}{|\mathcal{B}_b|};\ \mathbf{s} = \mathbf{s} - \eta \cdot \frac{g\mathbf{s}}{|\mathcal{B}_b|};$
    end for
end while
return $\{\boldsymbol{\theta}, \boldsymbol{\psi}, \mathbf{s}\}$
```

4 Experiments

The objectives are to investigate the impact of sentiment on text-to-image retrieval and to assess the efficacy of sentiment-oriented metric learning framework via comparison with various cross-modal retrieval baselines.

4.1 Experimental Setup

Datasets. We conduct experiments on two datasets including *Visual Sentiment Ontology (VSO)* [6] and online reviews crawled from *Yelp.com.*

VSO dataset consists of adjective-noun pairs (ANP), e.g., *delicious drink* or *angry face*, associated with sentiment scores. Images are retrieved from Flickr when using these ANPs as queries. Firstly, sentiment is binarized based on the sign of the scores. Secondly, to reduce sentiment biases, we neutralize the queries by only using the nouns. Images from all ANPs belonging to the same noun are merged together. To remove the biases, we balance the number of images between two sentiments within each query via uniform sampling. These would then form $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i)$ triples in $\mathcal{T}$, which is randomly split into 5 folds for model cross-validation. Statistics of the VSO dataset after being processed is shown in Table 1. The numbers of triples are not identical as not all queries have divisible-by-5 number of triples. A small fraction of images appear in multiple queries, thus, the number of images is smaller than the number of triples.

Table 1: Data statistics

		Fold 1	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Total**
VSO	#images	29,745	30,033	29,654	30,039	29,614	149,085
	#triples	29,798	30,088	29,704	30,080	29,660	149,330

		BO	**CH**	**LA**	**NY**	**SF**	**Total**
Yelp	#images	19,054	19,054	19,054	19,054	19,054	95,270
	#triples	38,303	37,643	38,816	37,762	38,654	191,178

Yelp dataset consists of reviews of businesses in 5 US cities: Boston (BO), Chicago (CH), Los Angeles (LA), New York (NY), and San Francisco (SF). Each review has a rating, review text, and one or more images taken by the user. Sentiment is derived from the rating score, whereby ratings 1 and 2 are considered negative, ratings 4 and 5 are considered positive, while rating 3 is dropped as being ambiguous. Review text is split into shorter passages; each sentence is considered a text query. An image can be paired with multiple queries from the same review. To identify the best-matching text-image pairs, we rank the text queries based on cosine similarity of their TF-IDF vectors to that of the user-provided image caption, and consider up to 3 highest-ranked text queries to be relevant. These form the $(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i)$ triples in $\mathcal{T}$. To neutralize a text query $\mathbf{x}_i$, words strongly suggestive of sentiment (i.e., *objective_score* < 0.5 by SentiWordNet [4]) are replaced by a special token -*MSK*-. We balance the number of images between the two sentiments and across the cities via uniform sampling. Table 1 shows statistics of the Yelp dataset after being processed. The numbers of triples are not identical as not all queries have 3 matched images.

Evaluation Protocols. We adopt a similar test procedure as [18, 39]. In our case, we conduct 5-fold validation, where for the *Yelp* dataset, four cities are used for training and one city is used for testing. During the test phase, for each query we construct a sample of 1,000 images, which include the correct images as well as uniformly sampled images in the test set. For each experiment, we report average result across 10 independent runs as well as the standard deviation.

Comparative Methods. We compare the proposed methods SML$_{OPPO}$ and SML$_{FLEX}$ with the following approaches:

- *Random* is the simplest baseline without learning,
- *CCA* [14] is one of the strongest statistical methods for cross-modal retrieval, which learns linear projections from input features, i.e., average Word2vec embeddings for text query and ResNet-50 features for images,
- *DCCA* [3] is the most recent extension of CCA transforming the same input features using multilayer perceptrons (i.e., we follow the original architecture of MLP in the original work) to capture non-linear interactions,
- *ACMR* [41] is a competitive method for cross-modal retrieval based on adversarial learning, in which modality-invariant representation in the common space is achieved by confusing a modality discriminator. We use the same neural network architectures for ours and ACMR for parity.

For all methods, the size of latent space is set to $D = 300$. For models that use stochastic gradient optimization, their parameters are updated with Adam [19] adaptive rule, batch size of 256, and learning rate of 0.001. Upon grid search for regularization $\lambda \in \{1e^{-5}, 1e^{-4}, \ldots, 1e^{-1}\}$ and margins $\tau_* \in \{0.0, 0.1, \ldots, 1.0\}$, the best hyper-parameters are obtained with cross-validation.

Metrics. We employ three established ranking metrics to measure the retrieval performance of the compared methods.

- *Percentile Rank (PR)* measures how well the correct images are being ranked amongst the image population. $PR = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{|D_i|}\sum_{j\in D_i}\frac{rank_j}{M}\right)$, where D_i denotes the set correct images for the query i, $rank_j$ is the rank of image j by the model, and M is the total number of images being ranked.
- *Normalized Discounted Cumulative Gain (NDCG)* measures the quality of ranking. $NDCG = \frac{1}{N}\sum_{i=1}^{N}\frac{DCG_i}{idealDCG_i}$, where $DCG_i = \sum_{j\in D_i}\frac{1}{\log{(rank_j+1)}}$, is the gain of image i relative to its position in a ranked list, and $idealDCG_i$ is the best achievable DCG_i in which all the correct images are at the top.
- *Recall@K (R@K)* denotes the ratio of correct images in the top-K retrieved images to the total number of correct images. $R@K = \frac{1}{N}\sum_{i=1}^{N}\frac{\sum_{j\in D_i}\mathbb{1}[j\in L_i]}{|D_i|}$, where $\mathbb{1}[*]$ is the indicator function and L_i is the top-K retrieved images.

4.2 Quantitative Evaluation

Comparison among Baselines. For an overall sense of the retrieval accuracy, Table 2 and Table 3 report the results of comparative approaches on different metrics on the two datasets, respectively. Random is the ground-level reference for relative comparisons with other methods.

The statistical method CCA shows a competitive performance. Starting with pre-trained embeddings from Word2Vec and ResNet-50, it benefits from the richly-compressed features from those underlying models, even though the projections it learns on top of these features are linear. DCCA obtains better results, attributable to further adaptation by learning non-linear transformations optimized for the same CCA objective. Even so, the gap between CCA and DCCA seems to be close on VSO dataset as the text queries are simpler (single nouns).

Considered a strong method for cross-modal retrieval, ACMR outperforms DCCA across all metrics on VSO and also on Yelp except for *Recall@10*. However, by adopting adversarial learning with less stable optimization [20], the variances of ACMR tend to be higher than other methods. That explains why DCCA can surpass ACMR on *Yelp-Recall@10*, which takes into account only *top-10* items rather than global ranking by *Percentile Rank* and *NDCG*.

Effect of Proposed Sentiment-Orientation. By leveraging sentiment information, both SML$_{OPPO}$ and SML$_{FLEX}$ significantly outperform all the sentiment-insensitive baselines across virtually all metrics and datasets. On average, SML$_{FLEX}$ model is slightly better than SML$_{OPPO}$. This is not unexpected as SML$_{OPPO}$ makes a stricter assumption on the direction of sentiment vectors.

Table 2: Performance of comparative methods on VSO dataset.

	Method	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Avg.
PR	**Random**	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00
	CCA	80.94 ± 0.00	81.05 ± 0.01	80.82 ± 0.01	80.67 ± 0.01	80.86 ± 0.01	80.87 ± 0.00
	DCCA	81.22 ± 0.11	81.33 ± 0.11	81.14 ± 0.05	81.00 ± 0.10	81.06 ± 0.09	81.15 ± 0.04
	ACMR	84.35 ± 0.21	84.43 ± 0.22	84.01 ± 0.14	84.16 ± 0.22	84.02 ± 0.26	84.19 ± 0.08
	SML$_{OPPO}$	**85.38** ± 0.10†	**85.42** ± 0.08†	85.04 ± 0.12†	**85.15** ± 0.06†	85.11 ± 0.11†	85.22 ± 0.04†
	SML$_{FLEX}$	85.34 ± 0.08†	**85.42** ± 0.11†	**85.10** ± 0.07†	85.14 ± 0.09†	**85.13** ± 0.07†	**85.23** ± 0.03†
NDCG (%)	**Random**	12.30 ± 0.02	12.32 ± 0.03	12.32 ± 0.03	12.31 ± 0.03	12.32 ± 0.03	12.31 ± 0.01
	CCA	19.55 ± 0.02	19.70 ± 0.02	19.59 ± 0.03	19.55 ± 0.04	19.62 ± 0.02	19.60 ± 0.01
	DCCA	20.08 ± 0.07	20.20 ± 0.10	19.96 ± 0.05	20.04 ± 0.07	20.06 ± 0.08	20.07 ± 0.03
	ACMR	20.64 ± 0.22	20.67 ± 0.20	20.41 ± 0.11	20.61 ± 0.21	20.56 ± 0.25	20.58 ± 0.09
	SML$_{OPPO}$	**21.95** ± 0.14†	21.93 ± 0.14†	21.74 ± 0.19†	21.80 ± 0.13†	21.89 ± 0.15†	21.86 ± 0.05†
	SML$_{FLEX}$	21.93 ± 0.15†	**21.97** ± 0.15†	**21.84** ± 0.12†	**21.87** ± 0.09†	**21.92** ± 0.13†	**21.91** ± 0.05†
R@10 (%)	**Random**	0.99 ± 0.07	1.00 ± 0.08	1.02 ± 0.05	0.99 ± 0.09	0.99 ± 0.05	1.00 ± 0.02
	CCA	12.00 ± 0.06	12.26 ± 0.06	12.01 ± 0.06	11.89 ± 0.09	12.19 ± 0.08	12.07 ± 0.03
	DCCA	13.25 ± 0.20	13.52 ± 0.25	13.08 ± 0.16	13.17 ± 0.13	13.23 ± 0.18	13.25 ± 0.08
	ACMR	14.01 ± 0.45	13.98 ± 0.41	13.61 ± 0.22	13.98 ± 0.49	13.94 ± 0.49	13.91 ± 0.18
	SML$_{OPPO}$	**16.75** ± 0.29†	16.77 ± 0.27†	16.43 ± 0.41†	16.50 ± 0.25†	**16.72** ± 0.38†	16.63 ± 0.13†
	SML$_{FLEX}$	**16.75** ± 0.36†	**16.85** ± 0.31†	**16.54** ± 0.26†	**16.57** ± 0.18†	16.71 ± 0.25†	**16.68** ± 0.12†

† improvements of *SML* models over the second-best baseline are statistically significant (p-value < 0.01).

Table 3: Performance of comparative methods on Yelp dataset.

	Method	BO	CH	LA	NY	SF	Avg.
PR	**Random**	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00	50.00 ± 0.00
	CCA	69.45 ± 0.01	68.65 ± 0.00	68.59 ± 0.01	69.01 ± 0.00	69.25 ± 0.01	68.99 ± 0.00
	DCCA	79.22 ± 0.26	78.67 ± 0.24	78.79 ± 0.34	79.01 ± 0.27	78.44 ± 0.28	78.83 ± 0.19
	ACMR	83.76 ± 0.89	83.32 ± 0.96	83.63 ± 0.65	83.67 ± 0.53	83.12 ± 0.80	83.50 ± 0.36
	SML$_{OPPO}$	**85.51** ± 0.09†	**84.84** ± 0.12†	84.89 ± 0.17†	84.92 ± 0.14†	84.32 ± 0.24†	84.89 ± 0.10†
	SML$_{FLEX}$	85.48 ± 0.12†	84.81 ± 0.10†	**84.93** ± 0.17†	**84.96** ± 0.13†	**84.38** ± 0.12†	**84.91** ± 0.07†
NDCG (%)	**Random**	12.65 ± 0.03	12.76 ± 0.03	12.38 ± 0.03	12.41 ± 0.03	12.60 ± 0.02	12.56 ± 0.01
	CCA	19.82 ± 0.04	19.21 ± 0.02	18.80 ± 0.02	18.89 ± 0.02	18.97 ± 0.01	19.14 ± 0.01
	DCCA	21.06 ± 0.21	20.85 ± 0.20	20.38 ± 0.24	20.54 ± 0.21	20.40 ± 0.20	20.64 ± 0.14
	ACMR	20.88 ± 0.91	21.00 ± 0.92	20.29 ± 0.70	20.59 ± 0.54	21.01 ± 0.76	20.75 ± 0.38
	SML$_{OPPO}$	**22.83** ± 0.14†	22.51 ± 0.26†	21.66 ± 0.21†	21.95 ± 0.31†	22.20 ± 0.46†	22.23 ± 0.16†
	SML$_{FLEX}$	22.82 ± 0.09†	**22.57** ± 0.25†	**21.77** ± 0.33†	**22.10** ± 0.40†	**22.44** ± 0.19†	**22.34** ± 0.16†
R@10 (%)	**Random**	0.96 ± 0.06	1.02 ± 0.08	0.99 ± 0.06	0.98 ± 0.06	1.01 ± 0.04	0.99 ± 0.02
	CCA	12.78 ± 0.05	11.31 ± 0.05	11.80 ± 0.04	11.56 ± 0.04	11.55 ± 0.04	11.80 ± 0.02
	DCCA	14.75 ± 0.47	13.96 ± 0.43	14.39 ± 0.49	14.62 ± 0.52	13.77 ± 0.40	14.30 ± 0.31
	ACMR	13.43 ± 1.94	13.45 ± 1.86	13.19 ± 1.52	13.81 ± 1.19	14.25 ± 1.60	13.62 ± 0.81
	SML$_{OPPO}$	**17.49** ± 0.27†	16.44 ± 0.58†	16.04 ± 0.51†	16.59 ± 0.69†	16.65 ± 0.92†	16.64 ± 0.34†
	SML$_{FLEX}$	17.45 ± 0.22†	**16.62** ± 0.56†	**16.21** ± 0.67†	**16.93** ± 0.84†	**17.10** ± 0.40†	**16.86** ± 0.32†

† improvements of *SML* models over the second-best baseline are statistically significant (p-value < 0.01).

Fig. 2 visualizes the learned metric spaces of SML with four sample queries: "bill", "service", "drink", and "toilet", and their sentiment-infused queries, by projecting their vectors onto 2D using PCA [44]. For SML$_{OPPO}$, we observe opposing directions between positive and negative sentiments. For SML$_{FLEX}$, they are not directly opposing but still form obtuse angles. This indicates a strong contrast of the sentiment concepts captured by the models. In addition, with the relaxation, SML$_{FLEX}$ can pull "bill" and "service" together, i.e., they are considered closer semantically as compared to "drink" or "toilet". This could be an explanation for the higher accuracies exhibited by SML$_{FLEX}$.

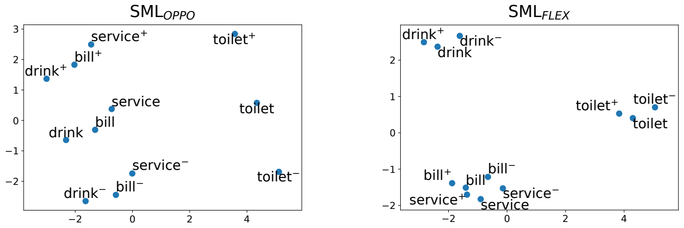

Fig. 2: Learned metric spaces of SML visualized in 2D using PCA.

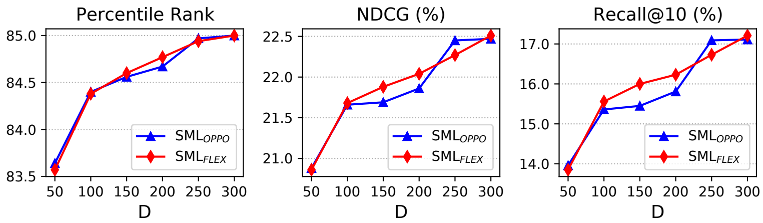

Fig. 3: Performance with varying the number of dimensions D of metric spaces.

Effect of Dimensionality. To further understand how the size of the metric space affects SML models, we conduct an experiment with different settings of dimensionality D on Yelp dataset. Fig. 3 illustrates performance of the SML$_{OPPO}$ and SML$_{FLEX}$ when D ranges from 50 to 300. Across all metrics, the model performances are sharply boosted when D increases from $50 - 200$ and tends to converge around the values of $250 - 300$, especially so in terms of *Percentile Rank*. Even though the performance of SML$_{FLEX}$ is potentially better if D goes beyond 300, it does not seem to be the case for SML$_{OPPO}$. Thus, we stop at $D = 300$, and all experiments are also conducted under this setting.

4.3 Case Studies

To gain more insights on the SML models, especially when the notion of sentiment is visually prominent, we illustrate examples from *Yelp-LA* dataset. Fig. 4 shows retrieved images with different queries and sentiments. In addition, we include ACMR as a reference baseline. In each ranking (top-4 are vertically positioned), the ground-truth is marked with a dotted rectangle. First of all, we notice that SML$_{FLEX}$ can retrieve the correct image in both cases and SML$_{OPPO}$ in one case. This observation concurs with the higher retrieval performance of SML$_{FLEX}$ in the previous quantitative analysis. Interestingly, in the second example, not only can SML$_{FLEX}$ pull the correct one into top-4, but it also illustrates a strong notion of sentiment when the first-ranked image, *"burned pizza"*, is evidently negative. Meanwhile, ACMR retrieves images based on the concepts implied by text queries, but not the ground-truth in both cases, presumably as it might not have captured the sentiment aspects well.

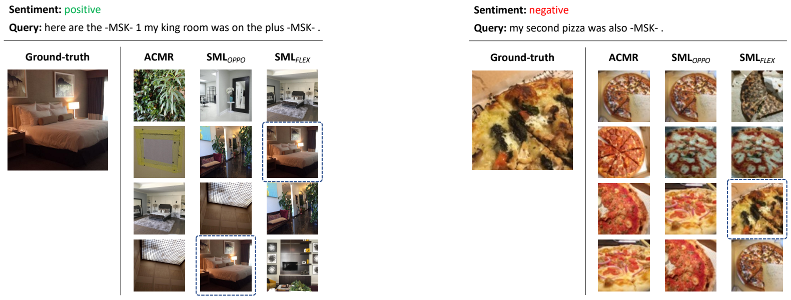

Fig. 4: Top retrieved images organized along queries.

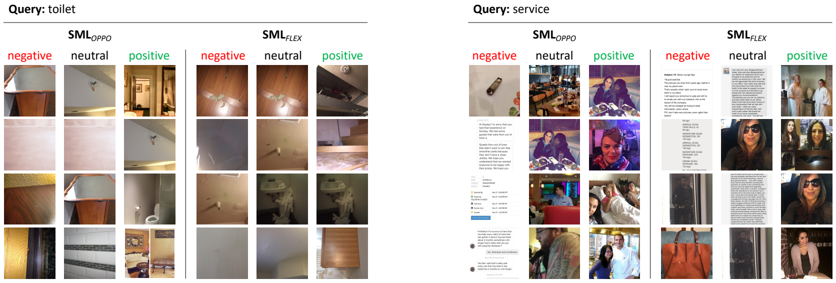

Fig. 5: Top retrieved images while changing sentiments.

For understanding the notion of sentiments captured by SML$_{OPPO}$ and SML$_{FLEX}$, in Fig. 5 we analyze 2 queries *"toilet"* and *"service"*, while alternating the sentiment input. Neutral means the sentiment vectors are set to zeros. For both queries, there are contrasts between *"negative"* and *"positive"* images. SML$_{OPPO}$ demonstrates that effect more clearly, especially on *"toilet"* query. This is due to desired constraint of the model, and can also be explained via Fig. 2 (i.e., sentiment vectors of *"toilet"* query are slightly longer in magnitude than the other queries'). For *"service"* query, negative images show complaint notes which imply customer unhappiness. Surprisingly, the positive images turn out to be smiling faces showing customer satisfaction. With such sentimental concepts captured via SML models, the case studies shed some light on understanding how the models work as well as how the performance could be interpreted.

5 Conclusion

We propose Sentiment-Oriented Metric Learning framework to incorporate sentiments into text-to-image retrieval. Our models SML$_{OPPO}$ and SML$_{FLEX}$ outperform comparable baselines on experiments involving images obtained from Flickr (VSO) as well as from online reviews (Yelp). As future work, the proposed framework could potentially be further extended to learn other visual concepts (e.g., human emotions, fashion styles) for text-to-image retrieval.

Acknowledgement

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its NRF Fellowship Programme (Award No. NRF-NRFF2016-07).

References

1. Akaho, S.: A kernel method for canonical correlation analysis. In: Proceedings of the International Meeting of the Psychometric Society (IMPS2001). Springer-Verlag (2001)
2. Anderson, T.: An introduction to multivariate statistical analysis.[una introducción al análisis estadístico multivariado] (1984)
3. Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis. In: ICML (2013)
4. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) LREC (2010)
5. Bach, F.R., Jordan, M.I.: Kernel independent component analysis. J. Mach. Learn. Res. **3**, 1–48 (2002)
6. Borth, D., Ji, R., Chen, T., Breuel, T., Chang, S.F.: Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: ACM Multimedia (2013)
7. Cao, Y., Long, M., Wang, J., Yang, Q., Yu, P.S.: Deep visual-semantic hashing for cross-modal retrieval. In: Krishnapuram, B., Shah, M., Smola, A.J., Aggarwal, C.C., Shen, D., Rastogi, R. (eds.) SIGKDD (2016)
8. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: ACM Multimedia (2014)
9. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: AISTATS (2011)
10. Gordo, A., Almazán, J., Revaud, J., Larlus, D.: Deep image retrieval: Learning global representations for image search. In: European conference on computer vision. pp. 241–257. Springer (2016)
11. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural computation **16**(12), 2639–2664 (2004)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
14. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
15. Hsieh, C., Yang, L., Cui, Y., Lin, T., Belongie, S.J., Estrin, D.: Collaborative metric learning. In: Barrett, R., Cummings, R., Agichtein, E., Gabrilovich, E. (eds.) WWW (2017)
16. Hsieh, W.W.: Nonlinear canonical correlation analysis by neural networks. Neural Networks **13**(10), 1095–1105 (2000)

17. Jiang, Q., Li, W.: Deep cross-modal hashing. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 3270–3278. IEEE Computer Society (2017)
18. Karpathy, A., Li, F.: Deep visual-semantic alignments for generating image descriptions. In: CVPR (2015)
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2015)
20. Kodali, N., Abernethy, J., Hays, J., Kira, Z.: On convergence and stability of gans. arXiv preprint arXiv:1705.07215 (2017)
21. Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P.M., Bischof, H.: Large scale metric learning from equivalence constraints. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2288–2295. IEEE (2012)
22. Kulis, B.: Metric learning: A survey. Foundations and Trends in Machine Learning **5**(4), 287–364 (2013)
23. Lai, P.L., Fyfe, C.: A neural implementation of canonical correlation analysis. Neural Networks **12**(10), 1391–1397 (1999)
24. Li, Z., Lin, D., Meng, H.M., Tang, X.: Discriminant mutual subspace learning for indoor and outdoor face recognition. In: 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA. IEEE Computer Society (2007)
25. Lin, D., Tang, X.: Inter-modality face recognition. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part IV. Lecture Notes in Computer Science, vol. 3954, pp. 13–26. Springer (2006)
26. Liu, W., Tsang, I.W.: Large margin metric learning for multi-label prediction. In: Bonet, B., Koenig, S. (eds.) Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. pp. 2800–2806. AAAI Press (2015)
27. Melzer, T., Reiter, M., Bischof, H.: Nonlinear feature extraction using generalized canonical correlation analysis. In: International Conference on Artificial Neural Networks. pp. 353–360. Springer (2001)
28. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Burges, C.J.C., Bottou, L., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. pp. 3111–3119 (2013)
29. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Getoor, L., Scheffer, T. (eds.) Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011. pp. 689–696. Omnipress (2011)
30. Peng, Y., Qi, J.: Cm-gans: Cross-modal generative adversarial networks for common representation learning. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) **15**(1), 1–24 (2019)
31. Ragusa, E., Cambria, E., Zunino, R., Gastaldo, P.: A survey on deep learning in image polarity detection: Balancing generalization performances and computational costs. Electronics **8**(7), 783 (2019)
32. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2160–2167. IEEE (2012)
33. Shen, F., Zhou, X., Yang, Y., Song, J., Shen, H.T., Tao, D.: A fast optimization method for general binary code learning. IEEE Transactions on Image Processing **25**(12), 5610–5621 (2016)

34. Song, J., Yang, Y., Yang, Y., Huang, Z., Shen, H.T.: Inter-media hashing for large-scale retrieval from heterogeneous data sources. In: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data. pp. 785–796 (2013)
35. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
36. Truong, Q.T., Lauw, H.W.: Visual sentiment analysis for review images with item-oriented and user-oriented cnn. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1274–1282 (2017)
37. Truong, Q.T., Lauw, H.W., Aumüller, M., Nitta, N.: Reproducibility Companion Paper: Visual Sentiment Analysis for Review Images with Item-Oriented and User-Oriented CNN, p. 4444–4447 (2020)
38. Vadicamo, L., Carrara, F., Cimino, A., Cresci, S., Dell'Orletta, F., Falchi, F., Tesconi, M.: Cross-media learning for image sentiment analysis in the wild. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 308–317 (2017)
39. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
40. Wan, J., Wang, D., Hoi, S.C.H., Wu, P., Zhu, J., Zhang, Y., Li, J.: Deep learning for content-based image retrieval: A comprehensive study. In: Proceedings of the 22nd ACM international conference on Multimedia. pp. 157–166 (2014)
41. Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 154–162 (2017)
42. Wang, J., He, Y., Kang, C., Xiang, S., Pan, C.: Image-text cross-modal retrieval via modality-specific feature learning. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval. pp. 347–354 (2015)
43. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: International Conference on Machine Learning. pp. 1083–1092 (2015)
44. Wold, S., Esbensen, K., Geladi, P.: Principal component analysis. Chemometrics and intelligent laboratory systems **2**(1-3), 37–52 (1987)
45. Xu, X., Shen, F., Yang, Y., Shen, H.T., Li, X.: Learning discriminative binary codes for large-scale cross-modal retrieval. IEEE Trans. Image Process. **26**(5), 2494–2507 (2017)
46. Xu, Z.E., Chen, M., Weinberger, K.Q., Sha, F.: From sbow to dcot marginalized encoders for text representation. In: Chen, X., Lebanon, G., Wang, H., Zaki, M.J. (eds.) 21st ACM International Conference on Information and Knowledge Management. pp. 1879–1884. ACM (2012)
47. You, Q., Luo, J., Jin, H., Yang, J.: Robust image sentiment analysis using progressively trained and domain transferred deep networks. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA. pp. 381–388. AAAI Press (2015)
48. Zhai, D., Chang, H., Shan, S., Chen, X., Gao, W.: Multiview metric learning with global consistency and local smoothness. ACM Trans. Intell. Syst. Technol. **3**(3), 53:1–53:22 (2012)
49. Zhai, X., Peng, Y., Xiao, J.: Heterogeneous metric learning with joint graph regularization for cross-media retrieval. In: Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. AAAI Press (2013)
50. Zheng, F., Tang, Y., Shao, L.: Hetero-manifold regularisation for cross-modal hashing. IEEE Trans. Pattern Anal. Mach. Intell. **40**(5), 1059–1071 (2018)