applied sciences

MDPI

Article

Soft Contrastive Cross-Modal Retrieval

Jiayu Song [1], Yuxuan Hu [1,*], Lei Zhu [2], Chengyuan Zhang [3], Jian Zhang [1] and Shichao Zhang [1]

[1] School of Computer Science and Engineering, Central South University, Changsha 410083, China; jiayusong@csu.edu.cn (J.S.); jianzhang@csu.edu.cn (J.Z.); zhangsc@csu.edu.cn (S.Z.)
[2] College of Information and Intelligence, Hunan Agricultural University, Changsha 410128, China; leizhu@hunau.edu.cn
[3] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China; cyzhangcse@hnu.edu.cn
* Correspondence: hyx_08@csu.edu.cn

Abstract: Cross-modal retrieval plays a key role in the Natural Language Processing area, which aims to retrieve one modality to another efficiently. Despite the notable achievements of existing cross-modal retrieval methodologies, the complexity of the embedding space increases with more complex models, leading to less interpretable and potentially overfitting representations. Most existing methods realize outstanding results based on datasets without any error or noise, but that is extremely ideal and leads to trained models lacking robustness. To solve these problems, in this paper, we propose a novel approach, Soft Contrastive Cross-Modal Retrieval (SCCMR), which integrates the deep cross-modal model with soft contrastive learning and smooth label cross-entropy learning to boost common subspace embedding and improve the generalizability and robustness of the model. To confirm the performance and effectiveness of SCCMR, we conduct extensive experiments comparing 12 state-of-the-art methods on three multi-modal datasets by using image–text retrieval as a showcase. The experimental results show that our proposed method outperforms the baselines.

Keywords: cross-modal retrieval; soft contrastive learning; smooth label learning; common subspace; deep learning

check for updates

Citation: Song, J.; Hu, Y.; Zhu, L.; Zhang, C.; Zhang, J.; Zhang, S. Soft Contrastive Cross-Modal Retrieval. *Appl. Sci.* **2024**, *14*, 1944. https://doi.org/10.3390/app14051944

Academic Editors: Andrea Prati and Chilukuri K. Mohan

Received: 6 December 2023
Revised: 10 January 2024
Accepted: 23 February 2024
Published: 27 February 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cross-modal retrieval [1,2] is a hot research spot in Natural Language Processing, especially with the explosive increase in multimedia data (text, image, video, and audio) and application procedures (e.g., Tiktok, Twitter, and YouTube). It intends to retrieve one modality based on another modality, which will support users to acquire various types of information in different categories of multimedia data. The top challenge of cross-modal retrieval is how to effectively embed multimodal representation [3–5] in one share space and measure the similarities of multimodal data, which aims to bridge the heterogeneity gap in general.

Based on the existing approaches [6–8], a common method to diminish the heterogeneity gap between different types of data and bridge the relation of objects is to construct a common embedding subspace [9–12], where the multimedia data can be transformed into the same dimensional representation. In this way, the similarity of samples can be calculated directly and accurately. The traditional methods utilize statistical analysis approaches, e.g., CCA [13] and MCCA [14] analyze cross-modal correlation or learn semantic representation by finding linear associations between different datasets, which are the theory foundation of most traditional models. However, these methods are restricted by the linear function of statistics because they cannot acquire non-linear semantic mapping. As deep neural networks (DNNs) [15–17] have grown by leaps and bounds in recent years, methods based on deep learning promote non-linear representation learning [18–21] under the strong power of DNNs.

Motivation. To address the current limitations in DNNs representation learning, particularly in the context of multimodal data processing, we revisit the conventional approach of leveraging DNNs for Natural Language Processing. Furthermore, that greatly weakens the generalizability of models and directly obstructs realistic application. Additionally, there is a prevailing and questionable assumption in recent research that datasets are free of noise, with image–text pairs being perfectly aligned and instance ground truths being completely accurate. Such an assumption is far from realistic in real-world scenarios, where noise and inaccuracies are often inherent in the data. These inaccuracies can lead to overfitting and poor model performance on unseen or noisy data, which is a common occurrence in practical applications.

Method. To handle this restriction of the current model, this article proposes a novel end-to-end framework, named Soft Contrastive Cross-Modal Retrieval (SCCMR), which integrates a deep cross-modal model with soft contrastive learning and smooth label learning to boost common subspace embedding and improve the generalizability and robustness of the model.

Contributions. The contributions of our work are shown as follows:

- We propose a novel end-to-end cross-modal retrieval model, termed the Soft Contrastive Cross-Modal Retrieval (SCCMR), which aims to combine soft contrastive learning and cross-modal learning. To the best of our knowledge, this is the first work to fuse soft contrastive learning into cross-modal retrieval to improve multimodal feature embedding.
- To solve the sharp and tortuous margin problem of feature embedding, we use soft contrastive learning for common subspace learning, which assists the model in finding the smooth boundary of embedding representation. To balance the noise data, we utilize smooth label cross-modal learning, which promotes the robustness of the model on real multimedia data.
- We carry out extensive experiments on three benchmark multimedia datasets, Wikipedia, NUS-WIDE, and Pascal Sentence. Results illuminate that our proposed method can outperform the current state-of-the-art methods in cross-modal retrieval, which demonstrates the excellent effectiveness of the proposed method.

Roadmap. The rest of this article is organized as follows. We review the existing approaches that are related to this paper on common space learning and deep hashing learning in Section 2. Section 3 discusses the proposed method SCCMR in detail. Section 4 presents the extensive experiments and the results of the performance evaluation. In the final part, Section 5 concludes this whole article.

2. Related Work

In this section, we briefly review the studies of cross-modal retrieval and contrastive learning.

2.1. Cross-Modal Retrieval

Cross-modal retrieval [22–24] is an important issue in the field of information retrieval and machine learning, as shown in Figure 1. It aims at retrieval or correlation matching between different types of data (such as images, text, audio, etc.), thereby achieving the effective interaction and correlation of multimodal data [25]. It has a wide range of needs in practical applications, such as multimedia retrieval [26], content recommendation systems, and intelligent search engines. How to diminish the heterogeneous gap and bridge the semantic gap between different modalities is a significant problem; hence, considerable researchers have proposed lots of methods from their unique perspectives [27]. Furthermore, we divide the existing methods into two categories: Traditional Cross-Modal Methods and Deep Learning-Based Methods.

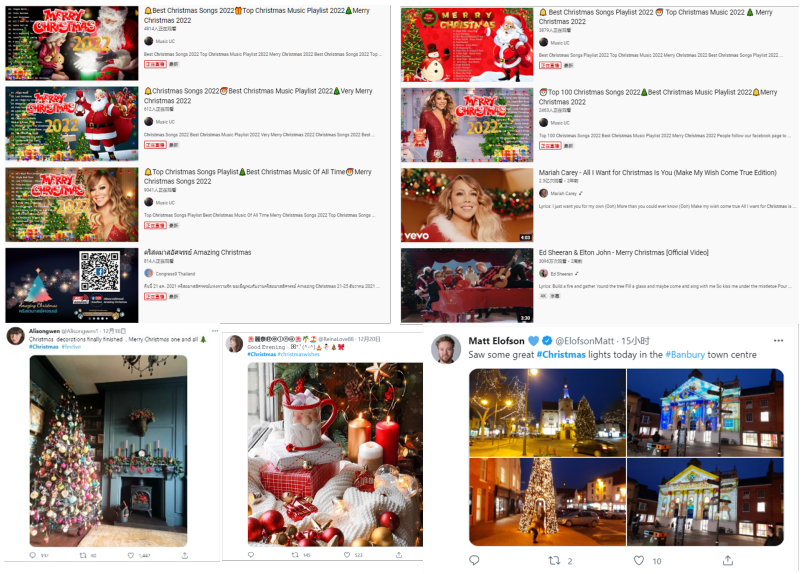


Figure 1. Multimedia data sample.

Traditional Cross-Modal Methods. Canonical correlation analysis (CCA) [13] stands as a classic multivariate method used to examine correlations between two sets of variables, or modalities. MCCA [14], building on CCA, seeks to uncover linear relationships and common structures across multiple datasets. Furthermore, Multiple View Discriminant Analysis (MvDA) [28] is designed to maximize the variability between different representations of the same object while minimizing variations within the same class. Kan et al. [28] applied MvDA to ensure consistent views across multiple linear transformations. Similarly, Zhai et al. [29] introduced an approach for joint representation learning to identify semantic connections and relevant features in cross-media data.

Deep Learning-Based Methods. Over the past two decades, deep learning has made significant advances in Natural Language Processing, with an increasing number of methods leveraging deep learning to address the challenge of semantic representation and bridge the heterogeneous gap [30]. For example, Zhu et al. [31] introduced a deep multigraph-based hierarchical enhanced semantic representation model (MG-HSF) to align semantic distributions at a fine-grained level. Wang et al. [32] proposed an adversarial cross-modal retrieval method (ACMR) that uses adversarial learning to discover a shared subspace. Wei et al. [33] developed a method for deep semantic matching to enhance the representational capabilities of CNNs. Peng et al. [34] suggested utilizing multiple deep networks to explore correlations in cross-media data through hierarchical learning. Conversely, Andrew et al. [35] presented deep canonical correlation analysis, which combines CCA with deep neural networks. Wang et al. [36] proposed a deep canonically correlated autoencoder approach for learning representations using DNNs.

2.2. Contrastive Learning

Contrastive learning [37] is a self-supervised learning technique designed to construct data representations that bring similar samples closer together in the representation space, while pushing dissimilar samples further apart. The fundamental concept involves learning feature representations [38] by optimizing the similarity of similar sample pairs and maximizing the distance between dissimilar sample pairs. This approach finds widespread use in representation learning for various types of data, including images, text, and audio. The typical approach involves constructing pairs of samples, namely positive and negative pairs, and then using a loss function to decrease the distance between similar samples while increasing the separation between dissimilar ones. The model then uses a loss function to encourage greater proximity between similar samples and increased separation between dissimilar ones. Through contrastive learning [39], models can acquire more discriminative

feature representations, which prove beneficial for various applications, including image retrieval [40–42], semantic matching [43,44], and unsupervised pre-training [45,46].

In the realm of traditional contrastive learning [47–49], Feng et al. [50] introduced adaptive soft contrastive learning, which reconceptualizes the task of sample discrimination into multi-instance soft discrimination. Zhang et al. [51] were the first to apply the soft contrastive loss to cross-modal problems, specifically in generative contexts. Liu et al. [52] proposed a unified model for multimodal retrieval, utilizing a singular contrastive training loss. This represents a significant advancement in cross-modal retrieval from the perspective of semantic embedding. However, while these approaches concentrate on aligning multimodal feature representations, they often overlook the impact of noisy data [53,54] and the challenges posed by uncertain labels.

3. Methodology

In this section, we present the proposed method SCCMR. Section 3.1 describes the problem definition. Section 3.2 presents the overview of the proposed framework. Sections 3.4–3.6 gives the method details. The summary of the notations is shown in Table 1.

3.1. Problem Definition

Definition 1. ***Cross-Modal Retrieval.*** *Given two modalities: an image modality I and a text modality T. $I = \{i_1, i_2, ..., i_n\}$ and $T = \{t_1, t_2, ..., t_m\}$ represent the set of images and text, where n and m are the number of images and texts, respectively. A cross-modal retrieval task can be defined as learning a similarity measure S, which can quantify the similarity between an image and a text:*

$$S : I \times T \rightarrow \mathbb{R} \tag{1}$$

Taking image-to-text retrieval as an example, given a query image i_q, the goal is to find the most matching text description t^* in the text set T:

$$t^* = \underset{t \in T}{\operatorname{argmax}}\, S(i_q, t) \tag{2}$$

Table 1. The summary of the notations frequently used in this paper.

Notation	Definition
$\mathcal{D}$	a multimedia dataset
$\boldsymbol{I}_i$	the i-th image sample
$\boldsymbol{T}_i$	the i-th text sample
$\boldsymbol{L}_i$	the semantic label vector of the i-th image–text pair
Q	a cross-modal query
$\mathcal{R}$	the set of results
N	the number of samples in a batch
$S(\cdot, \cdot)$	the multimodal semantic similarity function
$f_{\boldsymbol{I}}(\cdot)$	the mapping function of the visual network
$g_{\boldsymbol{T}}(\cdot)$	the mapping function of the textual network
$\theta_{\boldsymbol{I}}$	the parameters of the visual network
$\theta_{\boldsymbol{T}}$	the parameters of the textual network
τ	hyperparameter, temperature
ε	hyperparameter, smooth parameter
$\hat{p}(\cdot)$	the prediction value of the label classifier

Definition 2. ***Multi-Modal Similarity Function*** *Given a multi-modal dataset* $\mathcal{D}$, $(\boldsymbol{I}, \boldsymbol{T})$ *is a image–text pair, where* $\boldsymbol{I} \in \mathcal{D}$ *and* $\boldsymbol{T} \in \mathcal{D}$. *And the multi-modal similarity between* $\boldsymbol{I}$ *and* $\boldsymbol{T}$ *is calculated as follows, respectively:*

$$S(\boldsymbol{I}, \boldsymbol{T}) = \frac{\sum_i (f_I(I)_i \cdot g_T(T)_i)}{\sqrt{\sum_i (f_I(I)_i)^2} \cdot \sqrt{\sum_i (g_T(T)_i)^2}} \tag{3}$$

Here, f_I and g_T are the mapping functions of the visual network and textual network. For writing convenience, we write $S(f_I(\boldsymbol{I}), g_T(\boldsymbol{T}))$ as $S(\boldsymbol{I}, \boldsymbol{T})$ in the following part.

3.2. The Overview of SCCMR

Figure 2 demonstrates the overview of the SCCMR framework, which is an end-to-end cross-modal retrieval model via integrating soft contrastive learning and the smooth label cross-entropy learning method. This model consists of three main components: (1) cross-modal feature learning, (2) soft contrastive learning, and (3) smooth label cross-modal learning. The first module is to map images and texts into a common subspace via ImgNet and TextNet. The second module is a novel soft contrastive model, which aims at maximizing the similarity between positive samples while minimizing the similarity between negative samples in a common representation subspace. The third module consists of a label classifier with two fully connected layers and a smooth label cross-entropy loss that is significant for mining cross-modal semantic similarity and blurring noise labels to strengthen the generalizability of the model.

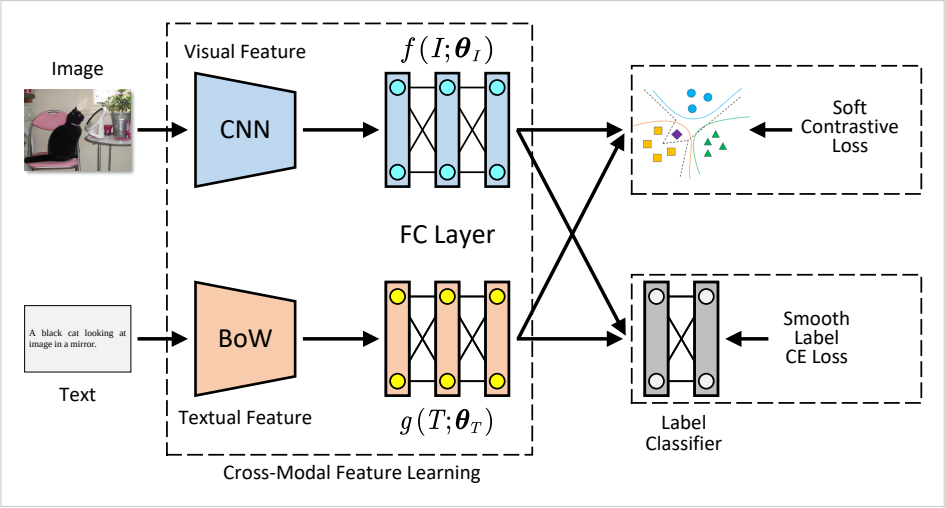


Figure 2. The framework of SCCMR.

3.3. Cross-Modal Feature Learning

For feature embedding learning, we employ two deep networks, ImgNet and TextNet, to embed each Image $I_i \in \mathcal{D}$ and text $T_i \in \mathcal{D}$ into a common representation subspace. For extracting high-level semantic features, ImgNet and TextNet both integrate several fully connected layers (FC layers).

ImgNet For images, each instance $\boldsymbol{I}_i$ is firstly encoded by a 19-layer VGG model [55], which is pre-trained on the ImageNet, and then fed into three FC layers to generate image feature embedding. In detail, we extract a 4096-dimension deep image feature vector by fc7 layer and conduct several fully connected layers ($\boldsymbol{I} \rightarrow 4096 \rightarrow 1000 \rightarrow 300$) activated by the *tanh* function to project image features $f(\boldsymbol{I}_i; \theta_{\boldsymbol{I}})$ into a common subspace.

TextNet For texts, we firstly utilize a widely used bag-of-word (BoW) model [56] to extract shallow textual features from each text $\boldsymbol{T}_j$, which is pre-trained in Google News, containing billions of words, and then use three FC layers activated by the *tanh* function ($\boldsymbol{T} \rightarrow 1000 \rightarrow 500 \rightarrow 300$) to output deep textual features $g(\boldsymbol{T}_j; \theta_{\boldsymbol{T}})$. The label classifier with

one FC layer realizes the classification of each modality ($f(\boldsymbol{I}) \setminus g(\boldsymbol{T}) \rightarrow c$), where c denotes the number of labels in multimodal dataset $\mathcal{D}$.

3.4. Soft Contrastive Learning

Contrastive learning [57] aims to associate different types of data (such as images and text) to achieve the goal of cross-modal retrieval, which is the mutual retrieval and matching of data in different modalities. Contrastive learning obtains a shared embedding space by learning the similarity between images and texts so that related images or texts are close to each other in the common space, while irrelevant samples are far away from each other. Undoubtedly, contrastive learning focuses on learning the common features between similar instances and distinguishing the differences between non-similar instances. Compared with generative learning, contrastive learning does not need to pay attention to the tedious details of the instance. It only needs to learn to distinguish the data in the feature space at the abstract semantic level. Consequently, the model and its optimization become simpler and have better generalization capabilities.

However, traditional contrastive learning utilizes the deep network model to pay much attention to noise samples or negative samples when overtraining, as a consequence, leading to the poor generalization ability of the model and poor retrieval results on new and unknown test data. This is because when overtraining or when the model is particularly complex and contains huge parameters, there will be sharp edges between different sample types in the common subspace as shown in Figure 3. During training, the mapping of the deep neural network makes each anomalous sample be divided into a specific boundary, resulting in an extremely tortuous and sharp classification boundary.

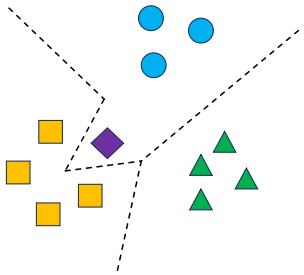


Figure 3. Illustration of the effect of overtraining or overfitting in feature embedding learning.

To better solve this problem, we use the soft contrastive loss to optimize the model parameters. In the process of sample mapping, the boundaries between classes become smooth. Despite it being a problem of model complexity or overtraining, soft contrast loss can help the model obtain smooth boundaries, thus greatly improving the generalization and robustness of the model as shown in Figure 4. Hence, the soft contrastive loss function is calculated by Equation (4) as follows:

$$\begin{aligned} \mathcal{L}_{scl} &= -\log\left(\frac{e^{\tau\cdot \mathrm{S}(\boldsymbol{I}_i,\boldsymbol{T}_j)}}{e^{\tau\cdot \mathrm{S}(\boldsymbol{I}_i,\boldsymbol{T}_j)} + \sum_{k=1}^{N} e^{\tau\cdot \mathrm{S}(\boldsymbol{I}_i,\boldsymbol{T}_k)}}\right) \\ &= -\tau\cdot \mathrm{S}(\boldsymbol{I}_i,\boldsymbol{T}_j) + \log\left(e^{\tau\cdot \mathrm{S}(\boldsymbol{I}_i,\boldsymbol{T}_j)} + \sum_{k=1}^{N} e^{\tau\cdot \mathrm{S}(\boldsymbol{I}_i,\boldsymbol{T}_k)}\right) \end{aligned} \tag{4}$$

where τ is an adjustable temperature parameter, controlling the distribution range of similarity. Equally, $\boldsymbol{I}_i$ and $\boldsymbol{T}_j$ are the embedding of the image and text, $S(\boldsymbol{I}_i, \boldsymbol{T}_j)$ denotes the cosine similarity between the text and image, and N is the total number of samples in the batch. The proof is shown in Appendix A.

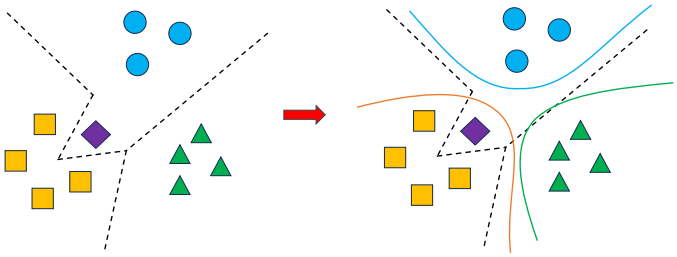


Figure 4. General idea of soft contrastive learning to improve sharp edges between classes.

3.5. Smooth Label Cross-Modal Learning

Even though there is a small amount of noise data [58] in existing datasets and the impact of uncertain labels on the retrieval results is minimal, considering the practical application of the model and for perfectly combining the certainty and uncertainty of data distribution, we introduced smooth label cross-entropy loss to balance the one-to-many problem and capacity caused by noise, which makes the model's prediction probability of training samples smoother by adding a smooth parameter to blur the label. Therefore, this reduces the model's absolute certainty about the label, thereby mitigating the overfitting problem. The smooth label cross-entropy loss is formulated by Equation (5) as follows:

$$\mathcal{L}_{sce} = -\sum_{i=1}^{n}((1-\varepsilon)\cdot y_i + \frac{\varepsilon}{n})(\log \hat{p}_i(\boldsymbol{I}_i) + \log \hat{p}_i(\boldsymbol{T}_i)) \tag{5}$$

where $\varepsilon \in (0,1)$ is a smooth parameter, y_i denotes the label of each instance, and n is the number of each mini-batch. $\hat{p}_i$ denotes the probability distribution of each modality, which is the prediction value of the label classifier.

Additionally, both soft contrastive loss and smooth label cross-entropy loss aim to promote the accuracy of retrieval results via weakening the effect of noise data and strengthening the generalizability of SCCMR. Under the result of the above discussion, the total objective loss function is described as Equation (6), respectively:

$$\mathcal{L}_{tot}(\theta_{\boldsymbol{I}}, \theta_{\boldsymbol{T}}) = \alpha\mathcal{L}_{scl} + \beta\mathcal{L}_{sce} \tag{6}$$

where the hyperparameters α and β decide the respective contributions of $\mathcal{L}_{scl}$ and $\mathcal{L}_{sce}$.

3.6. Optimization

We conduct the model in the process of learning the optimal function by minimizing the soft contrastive loss $\mathcal{L}_{scl}$ and smooth label cross-entropy loss $\mathcal{L}_{sce}$ as obtained in Equations (4) and (5), where $\hat{\theta_{\boldsymbol{I}}}$ and $\hat{\theta_{\boldsymbol{T}}}$ are the optimized parameters of visual network $f_{\boldsymbol{I}}$ and textual network $g_{\boldsymbol{T}}$, mathematically. As the optimization process of two networks is carried out simultaneously, the process runs as:

$$(\hat{\theta_{\boldsymbol{I}}}, \hat{\theta_{\boldsymbol{T}}}) = \arg\min_{\theta_{\boldsymbol{I}}, \theta_{\boldsymbol{T}}} \alpha\mathcal{L}_{scl} + \beta\mathcal{L}_{sce} \tag{7}$$

In addition, the visual network $f_{\boldsymbol{I}}$ is optimized by Adaptive Moment Estimation (Adam) in the following as Equation (8):

$$\theta_{\boldsymbol{I}} \longleftarrow \theta_{\boldsymbol{I}} - \lambda \cdot \nabla_{\theta_{\boldsymbol{I}}} \frac{1}{m}(\alpha\mathcal{L}_{scl} + \beta\mathcal{L}_{sce}) \tag{8}$$

Homosplastically, the textual network $g_{\boldsymbol{T}}$ is optimized by Adam as follows in Equation (9), respectively:

$$\theta_{\boldsymbol{T}} \longleftarrow \theta_{\boldsymbol{T}} - \lambda \cdot \nabla_{\theta_{\boldsymbol{T}}} \frac{1}{m}(\alpha\mathcal{L}_{scl} + \beta\mathcal{L}_{sce}) \tag{9}$$

Here, θ_I and θ_T are the parameters of visual network f_I and textual network g_T, λ denotes the learning rate, and m is the number of samples in mini-batch for each modality. Therefore, the complete optimization process is shown in Algorithm. 1.

Algorithm 1 Pseudocode of optimizing our SCCMR

Input: a training set $\mathcal{D} = \{\langle \boldsymbol{I}_i, \boldsymbol{T}_i, \boldsymbol{L}_i \rangle\}_{i=1}^{n}$, m samples in each mini-batch, the number of model training steps k, learning rate λ, hyperparameters α and β.
Initialize randomly θ_I, θ_T
repeat until convergence:
for k steps **do**
    Randomly select m image–text pairs from multimodal dataset $\mathcal{D}$ to construct a mini-batch$\{I_i, T_i\}$.
    Compute the representations $f_I(\boldsymbol{I})$ and $g_T(\boldsymbol{T})$ for the instance by forward propagation.
    Compute the soft contrastive loss $\mathcal{L}_{scl}$ and smooth label cross-entropy loss$\mathcal{L}_{sce}$ by Equations (4) and (5).
    Update the parameters θ_I of visual network f_I by Equation (8);
    Update the parameters θ_T of textual network g_T by Equation (9);
end for
Output: the optimized SCCMR model and learned representations in common subspace: $f_I(\boldsymbol{I})$ and $g_T(\boldsymbol{T})$

4. Experiment

In this section, we verify the remarkable performance of soft contrastive loss in cross-modal retrieval. We conduct comprehensive experiments on three benchmark datasets including Wikipedia, NUS-WIDE, and Pascal Sentence to evaluate the effectiveness of our proposed approach SCCMR compared with the state-of-the-art methods. Datasets and codes are available at https://github.com/Mario0716/SCCMR-master, accessed on 5 December 2023.

4.1. Setting

First of all, we introduce the settings of our experiments, including datasets, baselines, and evaluation metrics.

4.1.1. Datasets

We carry out all experiments on three popular datasets, Wikipedia [59], NUS-WIDE [60] and Pascal Sentence [61], which are introduced briefly as follows, and the statistic details of them are shown in Table 2. Besides, the image-text samples of datasets are shown in Figure 5.

Table 2. Scale and capacity of the three datasets used in our experiments.

Dataset	Train Set	Test Set	Labels	D_i	D_t
Wikipedia	1300	1566	10	4096	3000
NUS-WIDE	8000	1000	20	4096	1000
Pascal Sentence	800	100	20	4096	300

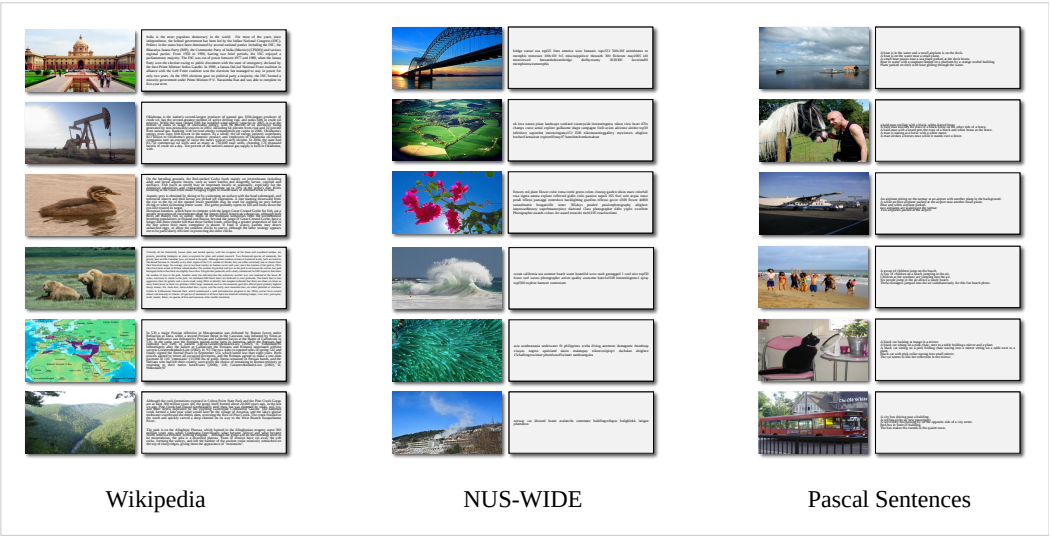


Figure 5. Sample image–text pairs of Wikipedia, NUS-WIDE, and Pascal Sentence.

Wikipedia. The Wikipedia dataset (https://en.wikipedia.org/wiki/, accessed on 5 December 2023) is composed of 2866 paired image–text documents extracted from "Wikipedia feature articles", which are considered to be some of the best content on Wikipedia. These documents are organized into ten semantic categories reflecting diverse interests: art and architecture; biology; geography and places; history; literature and theater; media; music; royalty and nobility; sport and recreation; and warfare. Each document includes a textual description that corresponds directly to the content of the associated image, providing a rich context for multimodal learning tasks. However, there are several notable challenges and potential biases inherent in this dataset: textual description quality, selection bias and representation bias.

NUS-WIDE. The NUS-WIDE dataset (https://lms.comp.nus.edu.sg/wp-content/uploads/2019/research/nuswide/NUS-WIDE.html, accessed on 5 December 2023) is a comprehensive collection of 269,648 images sourced from the Flickr platform. Each image of this dataset is accompanied by user-generated tags and cumulatively amounts to 5018 unique labels. This dataset, which serves as input for various image processing tasks, is characterized by its diverse visual content, representing a wide array of scenes, objects, and events. It includes extracted low-level features like color histograms, texture descriptors, and edge direction histograms. Additionally, the dataset provides ground-truth labels for 81 concepts out of the thousands of tags, which have been manually verified, facilitating precise evaluation of image annotation and retrieval algorithms. However, the NUS-WIDE dataset does present several challenges and potential biases: label noise, class imbalance, cultural bias and subjectivity in the ground truth.

Pascal Sentence. The Pascal Sentence Dataset (http://vision.cs.uiuc.edu/pascal-sentences/, accessed on 5 December 2023), a subset of the PASCAL Visual Object Classes (VOC), is designed to facilitate pattern analysis, statistical modeling, and computational learning. It encompasses twenty categories: aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, and TV. The training set comprises 10,103 images, featuring 23,374 objects, averaging approximately 500 objects per category. Some categories have imbalanced representation, which might lead to biases in model performance. For instance, 'person' is a category that is typically over-represented, which can result in models with a disproportionate sensitivity to human figures compared to other categories.

4.1.2. Baselines

To confirm the significant performance of our proposed method, we compare SC-CMR with 12 state-of-the-art methods in our experiments, including 6 traditional models

(CCA [13], SM [59], MCCA [14], MvDA [28], MvDA-VC [28] and JRL [29]) and 6 deep learning methods (Deep-SM [33], CMDN [34], DCCA [35], DCCAE [36], ACMR [32] and MG-HSF [31]). All of these baselines are introduced briefly as follows.

- **CCA** is a multivariate statistical method used to analyze the correlation between two sets of variables, often employed in cross-modal retrieval to learn correlations between different data modalities such as images and text.
- **SM** uses machine learning technology to represent cross-modal data as semantic vectors, facilitating retrieval and matching based on semantic content.
- **MCCA** is a statistical approach designed to explore relationships between multiple datasets, with the goal of finding linear associations that reveal shared structures or patterns.
- **MvDA** is an analysis method focused on maximizing the variance between different views of the same object or event while minimizing the variance within views of different objects or events.
- **MvDA-VC** is an extension of MvDA, which aims to maintain consistency across different views in the low-dimensional embedding space, thereby more effectively capturing the correlation and structural information between views.
- **JRL** is a method that improves information sharing and complementarity between different modalities or perspectives, leading to enhanced generalization ability and model performance.
- **Deep-SM** employs deep learning techniques for semantic matching, embedding data from different modalities like text and images into a neural network that facilitates cross-modal semantic similarity calculations.
- **CMDN** uses two multi-modal neural networks to embed and represent data, performing matching in a low-dimensional semantic space for cross-modal semantic similarity assessment.
- **DCCAE** consists of two main encoders and decoders that work together to encode input data into a low-dimensional representation and then attempt to reconstruct the original data, learning a compact representation for each modality.
- **ACMR** employs adversarial generative network techniques (specifically, Generative Adversarial Networks or GANs) to learn feature representations between modalities, using adversarial training to refine these representations.
- **MG-HSF** combines hierarchical semantic fusion with cross-modal adversarial learning to capture both fine-grained and coarse-grained semantic knowledge, generating modality-invariant representations in a common subspace.

4.1.3. Evaluation Metrics

mAP is a widely used evaluation metric in Natural Language Processing, especially in multi-modal retrieval. Its calculation involves computing the Average Precision (AP) for each query or class, and taking the mean across all classes or queries:

$$\boldsymbol{mAP} = \frac{1}{N}\sum_{k=1}^{N} AP_k \tag{10}$$

where N denotes the total number of queries, and AP_k is the Average Percision for k-th query.

4.1.4. Implementation Details

Our proposed method is implemented with the PyTorch 2.1.0 framework, on a workstation with Intel(R) CPU i9-12900KF 3.9 GHz, 64 GB RAM, and NVIDIA GEFORCE RTX 3090 with 24 GB memory. Furthermore, we optimize our proposed model SCCMR by Adam optimizer with a learning rate of 10^{-4} and set the maximum epoch as 500, τ as 0.7, and ε as 0.3.

4.2. Performance Evaluation

Secondly, we introduce the retrieval performance of our proposed method, the experimental analysis, and the ablation study.

4.2.1. Compared with State of the Art

In this part, we verify the effectiveness of SCCMR by comparing it with 12 baselines on three datasets: Wikipedia, NUS-WIDE, and Pascal Sentence.

Results on Wikipedia. We evaluate our proposed method by comparing it with 12 competitors on the Wikipedia dataset. Table 3 reports the mAP scores of the proposed SCCMR and the comparative methods. SCCMR outperforms both the traditional methods and deep learning methods by 61.90% and 49.23% in Text2Image and Image2Text tasks, respectively. Notably, SCCMR wins MG-HSF in Text2Image by mAP = 61.90%, even though MG-HSF defeats it in Image2Text by mAP = 52.85%. Specifically, our method is higher than MG-HSF in Text2Image by about 8.7%, which is a giant gap and significant improvement. Meanwhile, it verifies the great performance of SCCMR and proves that the combination of soft contrastive loss and smooth label cross-entropy loss achieves striking success in cross-modal retrieval. On the other hand, except for Deep-SM, DCCA, and DCCAE whose mAPs (Text2Image mAP = 35.43%, 39.60% and 38.50%; Image2Text mAP = 39.90%, 44.40%, and 43.50%) are a little lower than JRL (Text2Image mAP = 41.80%; Image2Text mAP = 44.90%), the traditional methods perform weaker than deep learning methods in general.

Table 3. The performance comparison (mAP@50 in %) with 12 competitors on the Wikipedia dataset. The highest performance values are shown in boldface.

Type	Method	Text2Image	Image2Text	Average
Traditional Method	CCA [13]	17.84	21.01	19.43
	SM [59]	28.51	23.34	25.93
	MCCA [14]	30.70	34.10	32.40
	MvDA [28]	30.80	33.70	32.30
	MvDA-VC [28]	35.80	38.80	37.30
	JRL [29]	41.80	44.90	43.40
Deep Learning Method	Deep-SM [33]	35.43	39.90	37.67
	CMDN [34]	42.70	48.70	45.70
	DCCA [35]	39.60	44.40	42.00
	DCCAE [36]	38.50	43.50	41.00
	ACMR [32]	43.42	47.74	45.58
	MG-HSF [31]	53.21	**52.85**	53.03
	Ours	**61.90**	49.23	**55.57**

The success of SCCMR can largely be credited to its innovative use of soft contrastive learning to discover a suitable common subspace for cross-modal retrieval. This technique allows for more nuanced and adaptable alignment of complex data structures, in contrast to the rigid semantic fusion networks employed by methods such as MG-HSF, which may struggle with the escalating complexity of data. The empirical evidence suggests that SCCMR not only improves the precision of cross-modal retrieval but also enhances the flexibility and scalability of the process, thus confirming its efficacy and potential for broader application in the field.

Results on NUS-WIDE. We testify our method by comparing it with 12 state-of-the-art methods on the NUS-WIDE dataset. The mAP scores of our proposed method SCCMR and other competitors on NUS-WIDE are shown in Table 4. Similar to those in the first group of comparisons, SCCMR outperforms all of its competitors by Text2Image mAP = 69.45% and Image2Text mAP = 67.97%, respectively. Unexpectedly, the precision of SCCMR is higher than the best competitor MG-HSF in Text2Image about 4.6% and Image2Text

indeed 6%, which is because MG-HSF pays too much attention to constructing semantic fusion networks to ensure absolute alignment. However, as the complexity of the data structure advances, this undoubtedly increases the difficulty of semantic alignment for each class, while SCCMR discovers a suitable common subspace by soft contrastive learning. Furthermore, the best performance and mAP gap compared with MG-HSF has ascertained the improvement and effectiveness of SCCMR in cross-modal retrieval. The disparity between SCCMR and MG-HSF can be attributed to their core methodological constructs. MG-HSF does not scale well with the increasing complexity of data structures, which intensively constructs semantic fusion networks to guarantee precise alignment. This complexity amplifies the difficulty of achieving semantic alignment across diverse classes, a challenge that the rigid MG-HSF alignment system may not effectively address.

Table 4. The performance comparison (mAP@50 in %) with 12 competitors on NUS-WIDE dataset. The highest performance values are shown in boldface.

Type	Method	Text2Image	Image2Text	Average
Traditional Method	CCA [13]	36.80	38.17	37.49
	SM [59]	42.37	39.16	40.77
	MCCA [14]	46.20	44.80	45.50
	MvDA [28]	52.60	50.10	51.30
	MvDA-VC [28]	55.70	52.60	54.20
	JRL [29]	59.80	58.60	59.20
Deep Learning Method	Deep-SM [33]	62.55	57.80	60.18
	CMDN [34]	51.50	49.20	50.40
	DCCA [35]	54.90	53.20	54.00
	DCCAE [36]	54.00	51.11	52.50
	ACMR [32]	57.85	58.41	58.13
	MG-HSF [31]	64.88	62.06	63.47
	Ours	**69.45**	**67.97**	**68.71**

Conversely, SCCMR introduces a soft contrastive learning paradigm that adeptly identifies a suitable common subspace. This flexible approach accommodates the intricate variances within the data, allowing enhanced semantic bridging between modalities without necessitating an absolute alignment of features. Hence, SCCMR not only ensures superior alignment between text and image data but also demonstrates adaptability to complex data structures. The results underscore the significant advancements SCCMR brings to cross-modal retrieval. The method's performance indicates its capability to generalize across modalities and effectively manage the multifaceted nature of the NUS-WIDE dataset. The mAP gap between SCCMR and MG-HSF solidifies the argument for the effectiveness and innovation of our approach. In conclusion, the pioneering SCCMR soft contrastive learning approach provides a robust solution for cross-modal retrieval tasks. The empirical results on the NUS-WIDE dataset validate its leadership position, opening avenues for further research and application in diverse scenarios requiring nuanced semantic understanding across different data types.

Results on Pascal Sentence. Our SCCMR method's performance has been rigorously evaluated against 12 state-of-the-art methods on the Pascal Sentence dataset as evidenced in Table 5. The table presents the mAP scores in percentages, with the highest values emphasized in bold. For the Text2Image retrieval, SCCMR achieves a mAP score of 69.55%, and that is significantly competitive, though it is not the highest, trailing by a narrow margin of 1.8% from the leading MG-HSF score of 71.55%. In the Image2Text retrieval task, our method secures a mAP score of 68.33%, which is only 1.29% shy of the leading MG-HSF score of 69.62%. The average performance of SCCMR is 68.94%, underscoring its robustness across both retrieval directions. When delving deeper into the comparative analysis, it is notable that SCCMR surpasses all traditional methods by a substantial

margin, indicating the advantages of leveraging deep learning for feature extraction and semantic understanding. Among traditional methods, MCCA shows the next best average performance of 67.70%, which our SCCMR method outperforms by 1.24%.

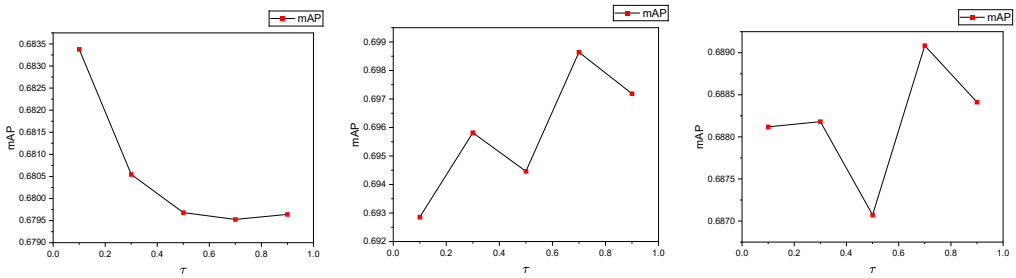
Among deep learning methods, the performance of our SCCMR is highly competitive. It is important to note that while MG-HSF leads in performance, the difference is narrow, and given the complexity of the data structure in the Pascal Sentence dataset, this gap indicates a significant achievement for SCCMR. The slight edge of MG-HSF can be attributed to its intense focus on constructing semantic fusion networks for alignment. However, our SCCMR method employs soft contrastive learning to find an optimal common subspace, which is particularly effective in handling the advanced complexity of semantic alignments across different classes. In summary, the close performance gaps between SCCMR and MG-HSF highlight the effectiveness of our proposed method. The soft contrastive learning approach of SCCMR proves to be a substantial contribution to this field, providing a robust alternative for cross-modal retrieval tasks. While MG-HSF emphasizes strict semantic alignments, the strategy of SCCMR to discover common subspaces accommodates the intricate structure of cross-modal data, enhancing retrieval precision.

Table 5. The performance comparison (mAP@50 in %) with 12 competitors on Pascal Sentence dataset. The highest performance values are shown in boldface.

Type	Method	Text2Image	Image2Text	Average
Traditional Method	CCA [13]	22.70	22.50	22.60
	SM [59]	21.12	18.74	20.14
	MCCA [14]	68.90	66.40	67.70
	MvDA [28]	62.60	59.40	61.00
	MvDA-VC [28]	67.30	64.80	66.10
	JRL [29]	53.40	52.70	53.10
Deep Learning Method	Deep-SM [33]	48.05	44.63	46.34
	CMDN [34]	52.60	54.40	53.50
	DCCA [35]	67.70	67.80	67.80
	DCCAE [36]	67.10	68.00	67.50
	ACMR [32]	67.60	67.10	67.35
	MG-HSF [31]	**71.55**	**69.62**	**70.59**
	Ours	69.55	68.33	68.94

4.2.2. Hyperparameter Sensitivity Analysis

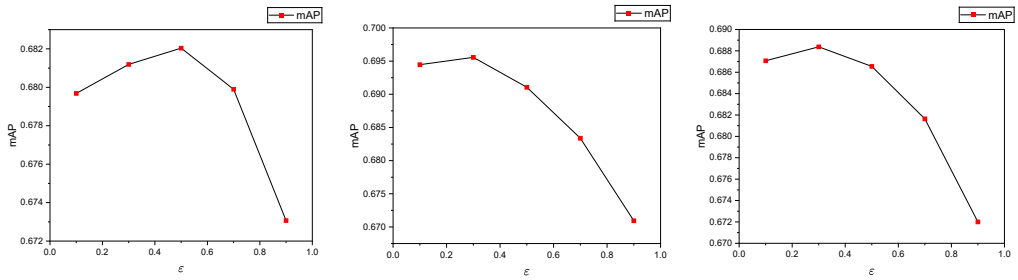
Temperature τ. Equation (4) illustrates the significant impact of the temperature parameter τ on the performance of SCCMR in cross-modal retrieval tasks. To corroborate this perspective, we performed a sensitivity analysis of the hyperparameter τ, utilizing the NUS-WIDE dataset across Image2Text, Text2Image, and average performance metrics. Figure 6 demonstrates that retrieval accuracy is subject to variation with adjustments in τ, with SCCMR realizing its highest mAP scores at a τ value of 0.7. These results endorse the premise that judicious selection of τ is pivotal for the optimal functioning of the SCCMR methodology.



(**a**) Image2Text (**b**) Text2Image (**c**) Average

Figure 6. Sensitivity analysis τ on NUS-WIDE dataset.

Hyperparameter ε. As shown in Equation (5), the hyperparameter ε notably affects the efficacy of SCCMR in cross-modal retrieval tasks. To substantiate this assertion, we examined the responsiveness of the hyperparameter ε within the NUS-WIDE datasets, considering Image2Text, Text2Image, and average performance metrics. As depicted in Figure 7, the retrieval accuracy exhibits variability with alterations in ε, with SCCMR reaching peak mAP when ε is configured to 0.3. This evidence corroborates the notion that an optimal ε selection is crucial for maximizing SCCMR performance.



(**a**) Image2Text (**b**) Text2Image (**c**) Average

Figure 7. Sensitivity analysis ε on NUS-WIDE dataset.

4.2.3. Ablation Study

In this section, we conduct the ablation study on the NUS-WIDE. As shown in Table 6, $\mathcal{L}_{soft}$ and $\mathcal{L}_{sce}$ are both essential to realize the best performance in cross-modal retrieval.

Table 6. Ablation study on NUS-WIDE with mAP@50 in %.

Method		Text2Image	Image2Text	Average
$\mathcal{L}_{scl}$	$\mathcal{L}_{sce}$			
	√	**69.52**	67.65	68.58
√		63.86	62.78	63.32
√	√	69.45	**67.97**	**68.71**

5. Conclusions

We proposed a novel end-to-end cross-modal retrieval method, named Soft Contrastive Cross-Modal Retrieval (SCCMR), which integrates soft contrastive learning and smooth label cross-modal learning for common embedding subspace. Specifically, the soft contrastive loss leverages a more flexible and forgiving approach to reducing the distance between similar pairs while maintaining a reasonable gap between dissimilar ones, effectively accommodating the subtle nuances of cross-modal data. Meanwhile, the smooth label cross-entropy loss aids in mitigating overfitting to noisy labels and encourages the model to consider label distribution, leading to a more balanced learning process. Thanks to the

soft contrastive loss and smooth label cross-entropy loss, we improve the generalizability and robustness of the retrieval model. In the experiments, we testify to the outstanding effectiveness and the demonstrated excellent performance compared with state-of-the-art methods on three benchmark datasets.

Author Contributions: Conceptualization, J.S.; methodology, J.S.; software, J.S.; validation, Y.H.; formal analysis, Y.H. and L.Z.; investigation, C.Z.; resources, J.S.; data curation, J.S.; original draft preparation, J.S.; supervision, J.Z. and S.Z.; project administration, J.S.; funding acquisition, C.Z., L.Z. and S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (62072166, 61836016, 61672177, 62202163), the Natural Science Foundation of Hunan Province (2023JJ30169, 2022JJ40190), the Scientific Research Project of Hunan Provincial Department of Education (22A0145).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All relevant data are within the paper.

Acknowledgments: We are grateful to the High Performance Computing Center of Central South University for assistance with the computations.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. SCL Loss Is Smoother than CL Loss

The form of traditional contrastive loss is given by:

$$L_{cl} = (1 - Y) \cdot D^2 + Y \cdot \max(0, m - D)^2 \tag{A1}$$

Here, m stands for margin, and D represents the distance between pairs of samples in the model's embedding space. For positive pairs ($Y = 1$) where $D < m$, we have:

$$\frac{\partial L_{cl}}{\partial D} = -2 \cdot (m - D), \quad \frac{\partial^2 L_{cl}}{\partial D^2} = 2 \tag{A2}$$

For negative pairs ($Y = 0$), the derivatives are:

$$\frac{\partial L_{cl}}{\partial D} = 2 \cdot D, \quad \frac{\partial^2 L_{cl}}{\partial D^2} = 2 \tag{A3}$$

The simplified form of the soft contrastive loss can be written as:

$$L_{scl} = -\log\left(\frac{\exp(-D)}{\exp(-D) + Z}\right) \tag{A4}$$

where Z is the sum of other terms. Note that for simplification, Z is assumed to be constant, although in an actual implementation of the soft contrastive loss, Z might depend on the sum of the distances of all negative pairs. The first derivative of the soft contrastive loss is:

$$\frac{\partial L_{scl}}{\partial D} = -\left(-\frac{\exp(-D)}{\exp(-D) + Z}\right) \cdot \left(1 - \frac{\exp(-D)}{\exp(-D) + Z}\right) \tag{A5}$$

And the second derivative is:

$$\frac{\partial^2 L_{scl}}{\partial D^2} = \frac{\exp(-D) \cdot (\exp(-D) + Z - \exp(-D))}{(\exp(-D) + Z)^2} \tag{A6}$$

Upon comparison, we observe that the second derivative of the soft contrastive loss contains a squared fraction term which changes smoothly with D, indicating lower sen-

sitivity to changes in the input, hence a smoother loss function. In contrast, the second derivative of the contrastive loss is a constant, indicating higher sensitivity to input changes, rendering the loss function less smooth than the soft contrastive loss.

References

1. Zhen, L.; Hu, P.; Wang, X.; Peng, D. Deep supervised cross-modal retrieval. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10394–10403.
2. Cao, Y.; Long, M.; Wang, J.; Yang, Q.; Yu, P.S. Deep visual-semantic hashing for cross-modal retrieval. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1445–1454.
3. Costa Pereira, J.; Coviello, E.; Doyle, G.; Rasiwasia, N.; Lanckriet, G.; Levy, R.; Vasconcelos, N. On the role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval. *Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 521–535. [CrossRef]
4. Chen, Y.; Yuan, J.; Tian, Y.; Geng, S.; Li, X.; Zhou, D.; Metaxas, D.N.; Yang, H. Revisiting multimodal representation in contrastive learning: From patch and token embeddings to finite discrete tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 15095–15104.
5. Lin, Z.; Bas, E.; Singh, K.Y.; Swaminathan, G.; Bhotika, R. Relaxing contrastiveness in multimodal representation learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 2227–2236.
6. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal learning with transformers: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef]
7. Fan, Y.; Xu, W.; Wang, H.; Wang, J.; Guo, S. PMR: Prototypical Modal Rebalance for Multimodal Learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 20029–20038.
8. Jin, P.; Huang, J.; Xiong, P.; Tian, S.; Liu, C.; Ji, X.; Yuan, L.; Chen, J. Video-text as game players: Hierarchical banzhaf interaction for cross-modal representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 2472–2482.
9. Rocco, I.; Arandjelović, R.; Sivic, J. End-to-End Weakly-Supervised Semantic Alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
10. Huang, S.; Wang, Q.; Zhang, S.; Yan, S.; He, X. Dynamic context correspondence network for semantic alignment. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2010–2019.
11. Hao, F.; He, F.; Cheng, J.; Wang, L.; Cao, J.; Tao, D. Collect and select: Semantic alignment metric learning for few-shot learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8460–8469.
12. Liu, Z.; Zhu, X.; Hu, G.; Guo, H.; Tang, M.; Lei, Z.; Robertson, N.M.; Wang, J. Semantic alignment: Finding semantically consistent ground-truth for facial landmark detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3467–3476.
13. Hotelling, H. Relations Between Two Sets of Variates. *Biometrika* **1936**, *28*, 321–377. [CrossRef]
14. Rupnik, J.; Shawe-Taylor, J. Multi-view canonical correlation analysis. In Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2010), Ljubljana, Slovenia, 12 October 2010; pp. 1–4.
15. Sze, V.; Chen, Y.H.; Yang, T.J.; Emer, J.S. Efficient processing of deep neural networks: A tutorial and survey. *Proc. IEEE* **2017**, *105*, 2295–2329. [CrossRef]
16. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
17. Marchetti, G.L.; Tegnér, G.; Varava, A.; Kragic, D. Equivariant representation learning via class-pose decomposition. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 25–27 April 2023; pp. 4745–4756.
18. Tao, C.; Zhu, X.; Su, W.; Huang, G.; Li, B.; Zhou, J.; Qiao, Y.; Wang, X.; Dai, J. Siamese image modeling for self-supervised vision representation learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada 18–22 June 2023; pp. 2132–2141.
19. Li, T.; Chang, H.; Mishra, S.; Zhang, H.; Katabi, D.; Krishnan, D. Mage: Masked generative encoder to unify representation learning and image synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada 18–22 June 2023; pp. 2142–2152.
20. Morioka, H.; Hyvarinen, A. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Valencia, Spain, 25–27 April 2023; pp. 3399–3426.
21. Cai, S.; Wang, Z.; Ma, X.; Liu, A.; Liang, Y. Open-world multi-task control through goal-aware representation learning and adaptive horizon prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada 18–22 June 2023; pp. 13734–13744.
22. Zhu, L.; Song, J.; Zhu, X.; Zhang, C.; Zhang, S.; Yuan, X. Adversarial learning-based semantic correlation representation for cross-modal retrieval. *IEEE MultiMedia* **2020**, *27*, 79–90. [CrossRef]

23. Zhu, L.; Zhang, C.; Song, J.; Liu, L.; Zhang, S.; Li, Y. Multi-graph based hierarchical semantic fusion for cross-modal representation. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 5–9 July 2021; pp. 1–6.
24. Jiang, D.; Ye, M. Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada 18–22 June 2023; pp. 2787–2797.
25. Liu, Y.; Li, G.; Lin, L. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 11624–11641. [CrossRef]
26. Chun, S.; Oh, S.J.; De Rezende, R.S.; Kalantidis, Y.; Larlus, D. Probabilistic embeddings for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8415–8424.
27. Song, Y.; Soleymani, M. Polysemous visual-semantic embedding for cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1979–1988.
28. Kan, M.; Shan, S.; Zhang, H.; Lao, S.; Chen, X. Multi-View Discriminant Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 188–194. [CrossRef]
29. Zhai, X.; Peng, Y.; Xiao, J. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *24*, 965–978. [CrossRef]
30. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
31. Zhu, L.; Zhang, C.; Song, J.; Zhang, S.; Tian, C.; Zhu, X. Deep multigraph hierarchical enhanced semantic representation for cross-modal retrieval. *IEEE MultiMedia* **2022**, *29*, 17–26. [CrossRef]
32. 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, Mountain View, CA, USA, 19–21 October 2017; pp. 154–162.
33. Wei, Y.; Zhao, Y.; Lu, C.; Wei, S.; Liu, L.; Zhu, Z.; Yan, S. Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Trans. Cybern.* **2016**, *47*, 449–460. [CrossRef]
34. Peng, Y.; Huang, X.; Qi, J. Cross-media shared representation by hierarchical learning with multiple deep networks. In Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI-16), New York, NY, USA, 9–15 July 2016; Volume 3846, p. 3853.
35. Andrew, G.; Arora, R.; Bilmes, J.; Livescu, K. Deep canonical correlation analysis. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 16–21 June 2013; pp. 1247–1255.
36. Wang, W.; Arora, R.; Livescu, K.; Bilmes, J. On deep multi-view representation learning. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 1083–1092.
37. Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; Krishnan, D. Supervised contrastive learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 18661–18673.
38. Le-Khac, P.H.; Healy, G.; Smeaton, A.F. Contrastive representation learning: A framework and review. *IEEE Access* **2020**, *8*, 193907–193934. [CrossRef]
39. Zang, Z.; Shang, L.; Yang, S.; Wang, F.; Sun, B.; Xie, X.; Li, S.Z. Boosting Novel Category Discovery Over Domains with Soft Contrastive Learning and All in One Classifier. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–3 October 2023; pp. 11858–11867.
40. Wang, X.; Han, X.; Huang, W.; Dong, D.; Scott, M.R. Multi-similarity loss with general pair weighting for deep metric learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5022–5030.
41. Cui, C.; Ma, Y.; Cao, X.; Ye, W.; Zhou, Y.; Liang, K.; Chen, J.; Lu, J.; Yang, Z.; Liao, K.D.; et al. A survey on multimodal large language models for autonomous driving. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 1–10 January 2024; pp. 958–979.
42. Sarukkai, V.; Li, L.; Ma, A.; Ré, C.; Fatahalian, K. Collage diffusion. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 1–10 January 2024; pp. 4208–4217.
43. Zhang, C.; Song, J.; Zhu, X.; Zhu, L.; Zhang, S. Hcmsl: Hybrid cross-modal similarity learning for cross-modal retrieval. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2021**, *17*, 1–22. [CrossRef]
44. Nie, Y.; Chen, H.; Bansal, M. Combining fact extraction and verification with neural semantic matching networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 6859–6866.
45. Barlow, H.B. Unsupervised learning. *Neural Comput.* **1989**, *1*, 295–311. [CrossRef]
46. Dy, J.G.; Brodley, C.E. Feature selection for unsupervised learning. *J. Mach. Learn. Res.* **2004**, *5*, 845–889.
47. Zolfaghari, M.; Zhu, Y.; Gehler, P.; Brox, T. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1450–1459.
48. Afham, M.; Dissanayake, I.; Dissanayake, D.; Dharmasiri, A.; Thilakarathna, K.; Rodrigo, R. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 9902–9912.
49. Kim, D.; Tsai, Y.H.; Zhuang, B.; Yu, X.; Sclaroff, S.; Saenko, K.; Chandraker, M. Learning cross-modal contrastive features for video domain adaptation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 13618–13627.

50. Feng, C.; Patras, I. Adaptive soft contrastive learning. In Proceedings of the 26th International Conference on Pattern Recognition (ICPR), Montreal, QC, Canada, 21–25 August 2022; pp. 2721–2727.
51. Zhang, H.; Koh, J.Y.; Baldridge, J.; Lee, H.; Yang, Y. Cross-modal contrastive learning for text-to-image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 833–842.
52. Liu, Z.; Xiong, C.; Lv, Y.; Liu, Z.; Yu, G. Universal Vision-Language Dense Retrieval: Learning A Unified Representation Space for Multi-Modal Retrieval. In Proceedings of the Eleventh International Conference on Learning Representations, Kigali Rwanda, 1–5 May 2022.
53. Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; Peng, X. Learning with noisy correspondence for cross-modal matching. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 29406–29419.
54. Natarajan, N.; Dhillon, I.S.; Ravikumar, P.K.; Tewari, A. Learning with noisy labels. *Adv. Neural Inf. Process. Syst.* **2013**, *26*.
55. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
56. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, *26*.
57. Tian, Y.; Sun, C.; Poole, B.; Krishnan, D.; Schmid, C.; Isola, P. What makes for good views for contrastive learning? *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6827–6839.
58. Müller, R.; Kornblith, S.; Hinton, G.E. When does label smoothing help? *Adv. Neural Inf. Process. Syst.* **2019**, *32*.
59. Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G.R.; Levy, R.; Vasconcelos, N. A new approach to cross-modal multimedia retrieval. In Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 25–29 October 2010; pp. 251–260.
60. Chua, T.S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; Zheng, Y. NUS-WIDE: A real-world web image database from National University of Singapore. In Proceedings of the ACM International Conference on Image and Video Retrieval. ACM, Santorini Island, Greece, 8–10 July 2009; p. 48.
61. Rashtchian, C.; Young, P.; Hodosh, M.; Hockenmaier, J. Collecting image annotations using amazon's mechanical turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, CA, USA, 6 June 2010; pp. 139–147.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.