Universal Weighting Metric Learning for Cross-Modal Matching

Jiwei Wei, Xing Xu*, Yang Yang, Yanli Ji, Zheng Wang, Heng Tao Shen
Center for Future Media & School of Computer Science and Engineering,
University of Electronic Science and Technology of China, China

Abstract

Cross-modal matching has been a highlighted research topic in both vision and language areas. Learning appropriate mining strategy to sample and weight informative pairs is crucial for the cross-modal matching performance. However, most existing metric learning methods are developed for unimodal matching, which is unsuitable for cross-modal matching on multimodal data with heterogeneous features. To address this problem, we propose a simple and interpretable universal weighting framework for cross-modal matching, which provides a tool to analyze the interpretability of various loss functions. Furthermore, we introduce a new polynomial loss under the universal weighting framework, which defines a weight function for the positive and negative informative pairs respectively. Experimental results on two image-text matching benchmarks and two video-text matching benchmarks validate the efficacy of the proposed method.

1. Introduction

Cross-modal matching aims at retrieving relevant instances of a different media type from the query, which has a variety of applications such as Image-Text matching [6, 32, 28, 36, 31, 37, 41, 2, 14], Video-Text matching [22, 29, 10, 38], Sketch-based image matching [3], etc. Compared with unimodal matching, cross-modal matching is more challenging due to the heterogeneous gap between different modalities. The key issue in cross-modal matching is to reduce the heterogeneous gap and exploit discriminative information across modalities [21, 8, 35, 33, 27].

A common solution for cross-modal matching is to learn a shared embedding space for different modalities so that the features from different modalities can be compared. Recently, a variety of cross-modal matching methods have been devoted to learning richer semantic representations for different modalities and a ranking loss is adopted to jointly optimize the network so that the similarity of the positive pairs is higher than that of all negative pairs, as illustrated in Figure 1. In previous literature, attention mechanism [18] and generative models [11, 15] have been explored to build advanced encoding networks. Liu *et al*. [23] proposed a recurrent residual fusion block to reduce the modality gap, and a triplet loss [13] was used to encourage semantically associated samples close to each other in the shared embedding space. Li *et al*. [19] proposed a visual reasoning model to generate global representation of a scene.

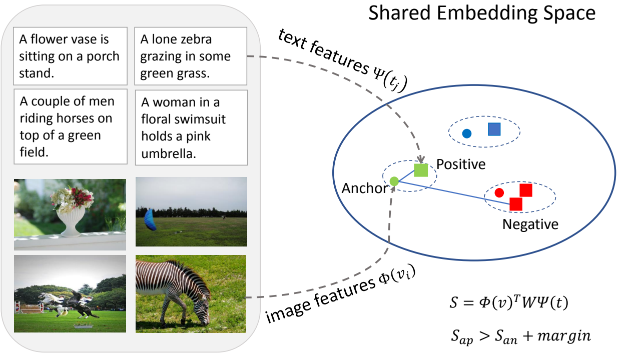

Shared Embedding Space
A flower vase is sitting on a porch stand.
A lone zebra grazing in some green grass.
A couple of men riding horses on top of a green field.
A woman in a floral swimsuit holds a pink umbrella.
text features $\Psi(t_j)$
image features $\Phi(v_i)$
Positive
Anchor
Negative
$S = \Phi(v)^T W \Psi(t)$
$S_{ap} > S_{an} + margin$

Figure 1. A typical solution for cross-modal matching is to learn a shared embedding space where visual features $\Phi(v)$ and text features $\Psi(t)$ can be compared. Points with the same shape are from the same modality. A triplet loss was utilized to encourage the similarity of positive (matching) pairs larger than negative (non-matching) pairs. Take image-text matching as an example.

While these methods have achieved encouraging performance, most of them use the ranking loss as an objective function, which usually trained with random sampling. This gives rise to an issue for cross-modal matching, where random sampling cannot effectively select informative pairs for training, leading a slow convergence and poor performance. While more recent metric learning methods have provided various mining strategies for unimodal matching, few of them are suitable for cross-modal matching. Hence, learning an appropriate mining strategy to sample and weight informative pairs is still a challenging problem for cross-modal matching.

*Corresponding author.

In this paper, we propose a universal weighting framework for cross-modal matching. Our intuition is based on the fact that a larger weight is assigned to a more informative pair, as illustrated in Figure 2. Unlike widely used unweighted triplet loss which treats all pairs equally, our proposed universal weighting framework can effectively assign appropriate weight to informative pairs for cross-modal matching. Specifically, we define two polynomial functions to calculate the weight values for positive and negative pairs respectively. Furthermore, we introduce a new polynomial loss under the universal weighting framework. Since the form of a polynomial function is flexible, our polynomial loss has a better generalization.

The major contributions of this paper are summarized as follows:

- We propose a universal weighting framework for cross-modal matching, which defines two polynomial functions to calculate the weight values for positive and negative pairs respectively. It provides a powerful tool to analyze the interpretability of various loss functions.
- We introduce a new polynomial loss under the universal weighting framework. The polynomial loss can effectively select informative pairs from redundant pairs, and assign appropriate weights to different pairs, resulting in performance boost.
- We conduct extensive experiments and evaluate our proposed method on two cross-modal matching tasks, image-text matching and video-text matching. Experimental results demonstrate that our method achieves very competitive performance on the four widely used benchmark datasets: MS-COCO, Flickr30K, ActivityNet-captions and MSR-VTT.

2. Related Work

Cross-Modal Matching. Cross-modal matching has a variety of applications, such as Image-Text matching [6, 32], Video-Text matching [9, 30, 22], Sketch-based image retrieval [3] etc. The key issue of cross-modal matching is measuring the similarity between different modal features. A common solution is to learn a shared embedding space where features of different modalities can be directly compared. In recent years, a variety of methods have been devoted to learning modality invariant features.

Lee *et al*. [18] proposed a stacked cross attention network for image-text matching, which measures the image-text similarity by aligning image regions and words. Li *et al*. [19] used graph convolutional network to generate relationship-enhanced image region features, then a global semantic reasoning network is performed to generate discriminative visual features that capture key objects and semantic concepts of a scene. Song *et al*. [30] introduced a polysemous instance embedding network that uses multi-head self-attention and residual learning to generate multiple representations of an instance. Liu *et al*. [22] proposed a collaborative expert (CE) framework for video-text matching, which generates dense representations for videos via aggregating information from different pre-trained models. Above embedding-based methods learn an advanced encoding network to generate richer semantic representations for different modalities, which make the matched pairs close to each other and the mismatched pairs far apart in the shared embedding space.

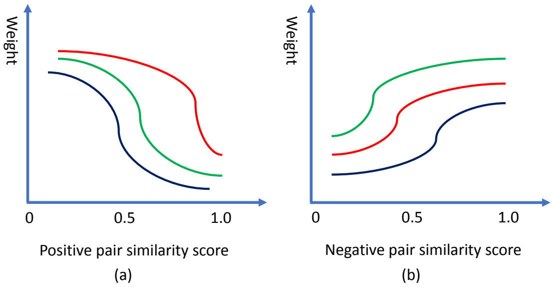


Figure 2. As the positive pairs similarity score increases, its weight value decreases; As the negative pairs similarity score increases, its weight value increases.

Metric Learning for Cross-Modal Matching. Another popular approach for cross-modal matching is to learn a loss function in the embedding space, which encourages the similarity of matched pairs larger than mismatched pairs. In recent years, a variety of metric learning methods have been proposed in both vision and language areas. However, most of existing metric learning methods are designed for uni-modal matching, which cannot effectively model the relationship of features captured from different modalities [21]. Only few of metric learning methods have been implemented particularly for cross-modal matching [35, 21, 6].

Liong *et al*. [21] introduced a deep coupled metric learning that designs two nonlinear transformations to reduce the modality map. Frome *et al*. [7] proposed a deep visual-semantic embedding model mapping visual features and semantic features into a shared embedding space, using a hinge rank loss as the objective function. Faghri *et al*. [6] introduced a variant triplet loss for image-text matching, and reported improved results. Xu *et al*. [35] introduced a modality classifier to ensure that the transformed features are statistically indistinguishable. However, these methods treat positive and negative pairs equally. Hardly any advanced sampling and weighting mechanism has been proposed for cross-modal matching. In this work, we present a universal weighting framework for cross-modal matching, which assigns a larger weight value to a harder sample.

3. The Proposed Approach

In this section, we formulate the sampling problem of cross-modal matching as a general weighting formulation. The proposed polynomial loss will be elaborated afterward.

3.1. Problem Statement

Let $v_i \in \mathbb{R}^{d_1}$ be a visual feature vector, $t_i \in \mathbb{R}^{d_2}$ be a text feature vector, $D = \{(v_i, t_i)\}$ be a training set of cross-modal instance pairs. In general, components of an instance pair come from different modalities. For simplicity, we refer to (v_i, t_i) as a positive pair and $(v_i, t_{j,i \neq j})$ as a negative pair. Given a query instance, the goal of cross-modal matching is to find a sample that matches it in another modality gallery. In the case of image-text matching, given an image caption t_i, the goal is to find the most relevant image v_i in the image gallery. It is important to note that in the cross-modal matching task, there is only one positive sample for each anchor in a mini-batch. Previous work for cross-modal matching focused on building a shared embedding space that contains both the image and text. The core idea behind these methods is that there exists a mapping function, $S(v, t; W) = \Phi(v)^T W \Psi(t)$ to measure the similarity score between the visual features $\Phi(v)$ and the text features $\Psi(t)$. W is the parameter of S. In general, the similarity score of the positive pair is higher than the negative pair by a margin, it can be formulated as:

$$S(v_i, t_i) > S(v_i, t_{j,j \neq i}) + \lambda_0, \forall v_i, \tag{1}$$

$$S(v_j, t_j) > S(v_{i,i \neq j}, t_j) + \lambda_0, \forall t_j, \tag{2}$$

where λ_0 is a fixed margin.

Since cross-modal matching is a mutual retrieval problem, the widely used triplet loss is formulated as :

$$L = [S(v, \hat{t}) - S(v, t) + \lambda_0]_+ + [S(\hat{v}, t) - S(v, t) + \lambda_0]_+, \tag{3}$$

where (v, t) is positive pair, $(v, \hat{t})$ is the hardest negative pair for a query v, and $(\hat{v}, t)$ is the hardest negative pair for a query caption t. $[x]_+ = max(x, 0)$. However, these methods discard pairs with less information than the hardest pair, while treating positive pairs and negative pairs equally. To our best knowledge, there is no advanced sampling and weighting method for cross-modal matching.

3.2. Universal Weighting Framework for Cross-Modal Matching

Let $N_{v_i} = \{S_{ij,i \neq j}\}$ be the set of similarity scores for all negative pairs of a sample v_i, and $N_{t_j} = \{S_{ij,j \neq i}\}$ be the set of similarity scores for all negative pairs of a sample t_j. Most existing hinge-based loss functions L can be formulated as a function of similarity scores: $L(\{S_{ij}\})$. Current existing weighting methods are given a special function to represent the relationship between weight values and similarity scores, the form of the function varies from task to task. All these functions can be reformulated into a universal weighting framework:

$$L = \sum_{i=1}^{i=N} \{G_{Pos} S_{ii} + \sum (G_{Neg} S_{ij,i \neq j})\}, \tag{4}$$

where G_{Pos} is the weight value of the positive pair, G_{Neg} is the weight value of negative pairs. Both of G_{Pos} and G_{Neg} are a function of similarity scores, but in a different forms.

$$G_{Pos} = G(S_{ii}, N_{v_i}), \tag{5}$$

$$G_{Neg} = G(S_{jj}, N_{t_j}), \tag{6}$$

where $G(\cdot)$ is a function that represents the relationship between weight value and similarity score. Theoretically, $G(\cdot)$ can be a function of self-similarity and relative similarity. The form of $G(\cdot)$ is various, but it should satisfy a basic rule: as the positive pairs' similarity score increases, its weight value decreases, and as the negative pairs' similarity score increases, its weight value increases. As illustrated in Figure 2. It provides a powerful tool to analyze the interpretability of various loss functions through weight analysis. Eq. 4 is a general pair formulation, existing pair-based loss is one of its special cases.

3.3. Informative Pairs Mining

For cross-modal matching tasks, in a mini-batch, each anchor has only one positive sample, but many negative pairs. These negative pairs are redundant and most of them are less informative. Random sampling are difficult to select more informative pairs, resulting in the model is difficult to convergence and have a poor performance. It is urgent and important to develop efficient algorithms that can select informative negative pairs and discard less informative negative pairs. In this section, we select informative negative pairs by comparing the relative similarity scores between positive and negative pairs of an anchor. For a given anchor v_i, we assume its positive sample is t_i, negative samples is $t_{j,i \neq j}$, and a negative pair (v_i, t_j) is selected if S_{ij} satisfies the condition:

$$S_{ij,i \neq j} > S_{ii} - \lambda, \tag{7}$$

where λ is a fixed margin. As illustrated in Figure 3. Note that there only one positive sample for each anchor in a mini-batch.

3.4. Polynomial Loss for Cross-Modal Matching

Through the above steps, negative pairs with more informative pairs can be selected and less informative pairs can be discarded. In this section, we introduce a new weighting function to weight the selected pairs. Theoretically,

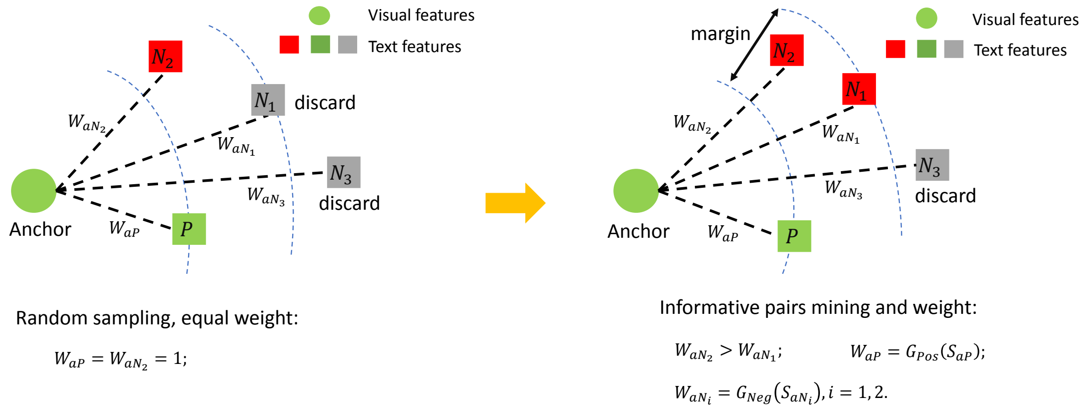


Figure 3. Illustration of our informative pairs mining and universal weighting framework for cross-modal matching. Points with the same shape are from the same modality. P is the only positive sample of anchor, N_1, N_2 and N_3 are the negative samples of anchor. Left: An example of random sampling and equal weighting; Right: The proposed negative pairs mining and universal weighting framework for cross-modal matching;

$G(\cdot)$ can be a function of self-similarity and relative similarity. However, the more complex the $G(\cdot)$, the more hyper-parameters it contains, and the hyper-parameters setting is more difficult. In this paper, to reduce the number of hyper-parameters, we define $G(\cdot)$ as a function of self-similarity. Specifically, given a selected positive pair (v_i, t_i), its weight G_{Pos} can be formulated as:

$$G_{Pos} = a_m S_{ii}^m + a_{m-1} S_{ii}^{m-1} + \cdots + a_1 S_{ii} + a_0, \quad (8)$$

where S_{ii} is the similarity score, $\{a_i\}_{i=0}^{i=m}$ is hyper-parameters, m is positive integer. The form of G_{Pos} is diverse, but its value should decrease with the increase of similarity score S_{ii}. Its trend should conform to the curve in figure 2a.

The weight G_{Neg} for a selected negative pair (v_i, t_j) can be formulated as:

$$G_{Neg} = b_k S_{ij}^k + b_{k-1} S_{ij}^{k-1} + \cdots + b_1 S_{ij} + b_0, i \neq j, \quad (9)$$

where S_{ij} is similarity score, $\{b_i\}_{i=0}^{i=k}$ is hyper-parameters, and k is positive integer. The form of G_{Neg} is diverse, but its trend should conform to the curve in figure 2b.

Through Eq. 8 and 9, we obtain the weights of positive and negative pairs. In this paper, we introduce two different functions, average polynomial loss and maximum polynomial loss.

Avg Polynomial Loss. Average polynomial loss can be defined as:

$$L_{Avg} = \frac{1}{N} \sum_{i=1}^{i=N} [G_{Pos} S_{ii}^p + \frac{\sum_{S_{ij} \in N_{v_i}} G_{Neg} S_{ij}^q}{Num(N_{v_i})} + \lambda_1]_+ + \frac{1}{N} \sum_{j=1}^{j=N} [G_{Pos} S_{jj}^p + \frac{\sum_{S_{ij} \in N_{t_j}} G_{Neg} S_{ij}^q}{Num(N_{t_j})} + \lambda_2]_+, \quad (10)$$

Eq. 10 can be reformulated as:

$$L_{Avg} = \frac{1}{N} \sum_{i=1}^{i=N} [\sum^{P} a_p S_{ii}^p + \frac{\sum_{S_{ij} \in N_{v_i}} \sum^{Q} b_q S_{ij}^q}{Num(N_{v_i})}]_+ + \frac{1}{N} \sum_{j=1}^{j=N} [\sum^{P} a_p S_{jj}^p + \frac{\sum_{S_{ij} \in N_{t_j}} \sum^{Q} b_q S_{ij}^q}{Num(N_{t_j})}]_+, \quad (11)$$

here, $Num(N_{v_i})$ and $Num(N_{t_j})$ denote the number of negative pairs of sample v_i and t_j respectively. P and Q are the highest power of positive and negative pairs, respectively. Note, we make the minimum of p and q to 0, and $a_0 = \lambda_1, b_0 = \lambda_2$.

For cross-modal matching tasks, only one positive sample for each anchor in a mini-batch. Our loss function can make full use of informative negative pairs. Since cross-modal matching tasks involve the mutual retrieval between different modalities, our loss comprises two terms. The former term represents the loss of image retrieval caption, and the latter represents the loss of caption retrieval image. Since the form of a polynomial function is flexible, our polynomial loss has a better generalization.

Max Polynomial Loss. To further highlight the superiority of our weighting mechanism, we introduce another version of polynomial loss L_{Max}, which only contains the hardest negative pair. The formulation of L_{Max} is defined as:

$$L_{Max} = \frac{1}{N} \sum_{i=1}^{i=N} [G_{Pos} S_{ii}^p + G_{Neg} Max\{N_{v_i}\}^q + \lambda_1]_+ + \frac{1}{N} \sum_{j=1}^{j=N} [G_{Pos} S_{jj}^p + G_{Neg} Max\{N_{t_j}\}^q + \lambda_2]_+, \quad (12)$$

here, $Max\{N_{v_i}\}$ and $Max\{N_{t_j}\}$ represent the hardest negative pair of sample v_i and t_j respectively. Eq.12 can be reformulated as:

$$L_{Max} = \frac{1}{N}\sum_{i=1}^{i=N}[\sum^{P} a_p S_{ii}^{p} + \sum^{Q} b_q Max\{N_{v_i}\}^{q}]_{+} + \frac{1}{N}\sum_{j=1}^{j=N}[\sum^{P} a_p S_{jj}^{p} + \sum^{Q} b_q Max\{N_{t_j}\}^{q}]_{+}, \tag{13}$$

Both L_{Avg} and L_{Max} can be minimized with gradient descent optimization. More discussions about L_{Avg} and L_{Max} can be found in the subsection of experiments.

4. Experiments

In this section, we conducted extensive experiments to evaluate the proposed polynomial loss in both image-text matching and video-text matching tasks. Following the [30, 18], we use the Recall@K as the performance metrics for both image-text matching and video-text matching tasks, which indicates the percentage of queries for which the model returns the correct item in its top K results. Ablation studies are conducted to analyze the effectiveness of proposed polynomial loss. We set the margin λ in Eq. 7 to 0.2 for all experiments.

4.1. Implementation Details

Image-Text Matching. We evaluate our polynomial loss on two standard benchmarks: MS-COCO [20] and Flickr30K [39]; **MS-COCO** dataset contains 123,287 images, and each image comes with 5 captions. We mirror the data split setting of [18]. More specifically, we use 113,287 images for training, 5,000 images for validation and 5,000 images for testing. We report results on both 1,000 test images (averaged over 5 folds) and full 5,000 test images. **Flickr30K** dataset contains 31,783 images, each image is annotated with 5 sentences. Following the data split of [18], we use 1,000 images for validation, 1,000 images for testing and the remaining for training.

Our implementation follows the practice in Stacked Cross Attention Network (SCAN) [18]. SCAN maps image regions and words into a shared embedding space to measure the similarity score between an image and a caption. For fair comparison, we keep the network structure unchanged and replace the loss function with polynomial loss. There are two inputs for SCAN, a set of image features which extracted by a pretrained Faster-RCNN model [1] with ResNet-101 [12], and a set of word features which encoded by a bi-directional Gated Recurrent Unit (GRU) [26]. Models are trained from scratch using Adam [16] with batch size of 128 for both datasets. For MS-COCO, we start training with learning rate 0.0005 for 10 epochs, and then lower it to 0.00005 for another 10 epochs. For Flickr30K, the learning rate is 0.0002 for 15 epochs, and then lower it to 0.00002 for another 15 epochs. There are two sets of parameters in the polynomial loss, $\{a_p\}$ and $\{b_q\}$. We adopt a heuristic method to select hyper-parameters. Concretely, we first initialize the $G(\cdot)$ to ensure that its curve conforms to the trend in Figure 2. Then, a grid search technology is adopted to select hyper-parameters. We set $P = 2$, $\{a_0 = 0.5, a_1 = -0.7, a_2 = 0.2\}$, $Q = 2$ and $\{b_0 = 0.03, b_1 = -0.3, b_2 = 1.2\}$ for MS-COCO and $P = 2$, $\{a_0 = 0.6, a_1 = -0.7, a_2 = 0.2\}$, $Q = 2$, $\{b_0 = 0.03, b_1 = -0.4, b_2 = 0.9\}$ for Flickr30K.

Video-Text Matching. We evaluate our polynomial loss on two popular datasets: ActivityNet-captions [17] and MSR-VTT [34]. **ActivityNet-captions** contains 20K videos, and each video comes with 5 text descriptions. We follow the data split of [22], 10,009 videos for training and 4,917 for testing. **MSR-VTT** contains 10K videos and each video is associated with about 20 sentences. We follow the data split of [22], 6,513 videos for training, and 2,990 videos for testing.

We report results on video-text matching task with Collaborative Experts (CE) [22] framework. CE is a framework that aggregated various pretrained features of a video into a dense representation before mapping to the shared embedding space. We keep the network structure unchanged and replace the loss function with polynomial loss. Models are trained from scratch using Adam [16] with batch size of 64 for both datasets. The learning rate is set to 0.0004. There are two sets of parameters in the polynomial loss, $\{a_p\}$ and $\{b_q\}$. We set $P = 2$, $\{a_0 = 0.5, a_1 = -0.7, a_2 = 0.2\}$, $Q = 2$ and $\{b_0 = 1, b_1 = -0.2, b_2 = 1.7\}$ for ActivityNet-captions and $P = 2$, $\{a_0 = 0.5, a_1 = -0.7, a_2 = 0.2\}$, $Q = 2$, $\{b_0 = 0.03, b_1 = -0.3, b_2 = 1.8\}$ for MSR-VTT.

4.2. Image-Text Matching Results

For image-text matching task, we compare the performance of our method with the several state-of-the-art methods, including: PVSE [30], VSE++ [6], SCO [15], RRF [23], DAN [25], GXN [11]and SCAN [18]. Table 1 and Table 2 summarize the results of our method on the Flickr30K and MS-COCO datasets, respectively. We also list the loss function used by various methods. From the table, we can make the following observations:

- From Table 1, we find the proposed method outperforms the baseline SCAN at all metrics. Compared with the classical triplet loss, the performance of SCAN with polynomial loss improves R@1 by 3.6% for text to image retrieval and 1.5% for image to text retrieval on Flickr30K.

- Table 2 summarizes the results on MS-COCO dataset. From Table 2, we can observe that the proposed

Methods	Loss Function	Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10
RRF [23]	Triplet	47.6	77.4	87.1	35.4	68.3	79.9
VSE++ [6]	Triplet	52.9	80.5	87.2	39.6	70.1	79.5
DAN [25]	Triplet	55.0	81.8	89.0	39.4	69.2	79.1
SCO [15]	Triplet+NLL	55.5	82.0	89.3	41.1	70.5	80.1
SCAN (I2T) [18]	Triplet	67.9	89.0	94.4	43.9	74.2	82.8
SCAN (I2T)	**Max Polynomial Loss**	**69.4**	**89.9**	**95.4**	**47.5**	**75.5**	**83.1**

Table 1. Experimental results on Flickr30K.

Methods	Loss Function	Image-to-Text			Text-to-Image		
		R@1	R@5	R@10	R@1	R@5	R@10
1K Test images							
VSE++ [6]	Triplet	64.6	89.1	95.7	52.0	83.1	92.0
GXN [11]	Triplet	68.5	-	97.9	56.6	-	**94.5**
PVSE [30]	Triplet+L_{div}+L_{mmd}	69.2	91.6	96.6	55.2	86.5	93.7
SCAN (I2T) [18]	Triplet	69.2	93.2	97.5	54.4	86.0	93.6
SCAN (I2T)	**Max Polynomial Loss**	**71.1**	**93.7**	**98.2**	**56.8**	**86.7**	93.0
5K Test images							
VSE++ [6]	Triplet	41.3	71.1	81.2	30.3	59.4	72.4
GXN [11]	Triplet	42.0	-	84.7	31.7	-	74.6
PVSE [30]	Triplet+L_{div}+L_{mmd}	45.2	74.3	84.5	32.4	63.0	75.0
SCAN (I2T) [18]	Triplet	46.4	77.4	87.2	34.4	63.7	75.7
SCAN (I2T)	**Max Polynomial Loss**	**46.9**	**77.7**	**87.6**	**34.4**	**64.2**	**75.9**

Table 2. Experimental results on MS-COCO.

method outperforms the state-of-the-art approaches, especially for R@1. By replacing the triplet loss with our L_{Max}, the performance of SCAN improves 1.9% on image to text retrieval (R@1) and 2.4% on text to image retrieval (R@1) on 1K test images.

- Classical triplet loss tries to sample the informative pairs from redundant pairs, but treats positive and negative pairs equally. In contrast to it, the proposed polynomial loss assigns appropriate weight value to the positive and negative pairs, and the weight value is related to its similarity score. The proposed method can simultaneously select and weight informative pairs. Extensive experimental results demonstrated that the proposed polynomial loss improved the matching performance effectively.

4.3. Video-Text Matching Results

We evaluate the effectiveness of our method on two standard benchmarks: ActivityNet-captions and MSR-VTT. We report our results and comparison with current state-of-the-art methods for video-to-text and text-to-video retrievals. The results are summarized in the Table 3 and Table 4 for ActivityNet-captions and MSR-VTT datasets, respectively.

In order to promote a comprehensive comparison, we list existing state-of-the-art results on these datasets, including: DENSE [17], HSE [40], CE [22] for ActivityNet-captions dataset and Minthum *et al.* [24], W2VV [4], CE [22] and Dual encoding [5] for MSR-VTT dataset. Furthermore, we list the loss function used by various methods. From Table 3 and Table 4, we can observe that our method outperforms the baselines on all measures, and achieves the new state-of-the-art performance on the video-text matching task. When compared with CE (Triplet) which uses the same video and sentence encoders with our method, our method improves 2.5% on text-to-video (R@1) task on the MSR-VTT dataset. Our method outperforms the CE on all metrics on the ActivityNet-captions dataset. The performance gap between CE (Triplet) and CE (Max Polynomial Loss) shows the effectiveness of our polynomial Loss.

4.4. Ablation Study

Parameter Analysis. There are two sets of parameters in the polynomial loss, $\{a_i\}$ and $\{b_j\}$. It is worth exploring to seek a set of parameters to make the model converge faster and achieves better performance. Since the number of hyper-parameters is too large, it is almost impossible to analyze the sensitivity of the hyper-parameters one by one, so we mainly analyze the sensitivity of several hyper-parameters with great influence. P and Q respectively determines the highest power of G_{Pos} and G_{Neg}, which has

Methods	Loss Function	Video-to-Text			Text-to-Video		
		R@1	R@5	R@10	R@1	R@5	R@10
DENSE [17]	Cross-entropy	18.0	36.0	74.0	14.0	32.0	65.0
HSE (4SEGS) [40]	Multi-loss	18.7	48.1	-	20.5	49.3	-
CE [22]	Triplet	27.9	61.6	**95.0**	27.3	61.1	94.4
CE	**Max Polynomial Loss**	**27.9**	**61.9**	94.1	**28.5**	**62.6**	**94.9**

Table 3. Experimental results on ActivityNet-captions.

Methods	Loss Function	Video-to-Text			Text-to-Video		
		R@1	R@5	R@10	R@1	R@5	R@10
Minthum wt al. [24]	Cross-entropy	12.5	32.1	42.4	7.0	20.9	29.7
W2VV [4]	Multi-loss	11.8	28.9	39.1	6.1	18.7	27.5
Dual encoding [5]	Triplet	13.0	30.8	43.3	7.7	22.0	31.8
CE [22]	Triplet	34.4	64.6	77.0	22.5	52.1	65.5
CE	**Max Polynomial Loss**	**36.2**	**71.5**	**82.2**	**25.0**	**55.4**	**68.2**

Table 4. Experimental results on MSR-VTT.

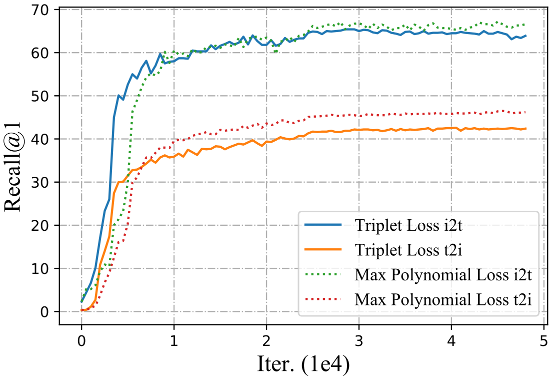


Figure 4. Triplet loss vs. Max Polynomial Loss on Flickr30K validation set. By replacing the loss function with our polynomial loss, the performance of SCAN is further improved.

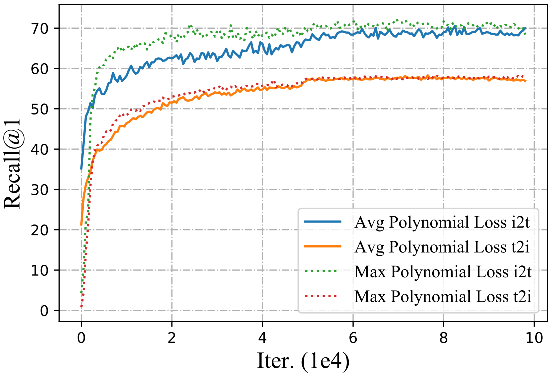


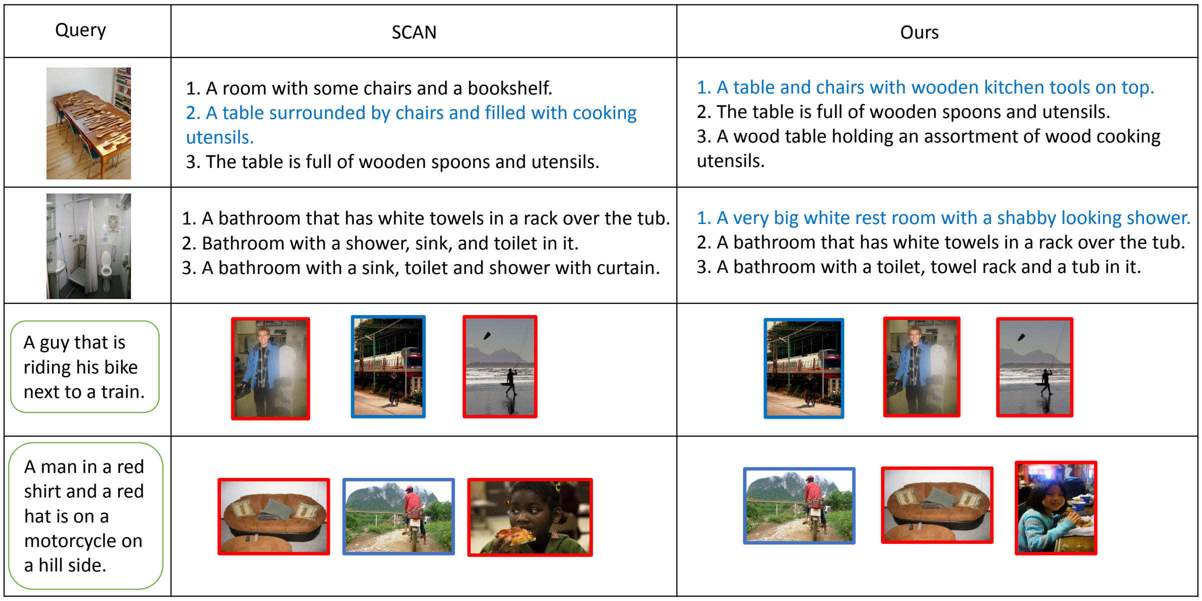
Figure 5. Analysis of the behaviors of the Max and Avg polynomial loss on MS-COCO validation set.

a direct impact on the number of hyper-parameters. Therefore, we first fixed them to 2. In practice, we find that model performance is most sensitive to parameters $\{b_q\}$, thus we mainly analyze the sensitivity of parameters $\{b_q\}$. We test the effect of b_1 and b_2 by fixing $b_0 = 0.03$, results are summarized in Table 5. b_1 and b_2 impact the hard level of negative pairs, the model is sensitive to different values. However, all of these combinations outperform the baseline, which demonstrates the superiority of our approach.

Triplet Loss *vs.* Max Polynomial Loss. Triplet loss is the most frequently used loss function for cross-modal matching tasks. Its effectiveness has been proved by many works, such as [6]. In this section, we further analyze the effectiveness of the proposed polynomial weighting mechanism. Max polynomial loss only includes the hardest negative pairs, which can be considered as a weighted version of triplet loss. We compare the max polynomial loss with triplet loss on MS-COCO dataset, the results are shown in Figure 4. From the results, we find the max polynomial loss converges faster than triplet loss and achieves a better result, which proven the superiority of our polynomial weighting mechanism.

Max *vs.* Avg Ploynomial Loss. In this section, we further analyze the effectiveness of the proposed Max and Avg polynomial loss. Max polynomial loss weights the positive pair and the hardest negative pair for each anchor, which can be considered as a weighted version of the hardest triplet loss. In contrast, average polynomial loss contains all of informative negative pairs and assigns different weight values for them. Since the max polynomial loss only utilizes a subset of informative pairs so that its computational complexity is lower than the average polynomial loss which contains all of informative negative pairs.

Figure 5 shows the performance of two functions on MS-



Query	SCAN	Ours
	1. A room with some chairs and a bookshelf. 2. A table surrounded by chairs and filled with cooking utensils. 3. The table is full of wooden spoons and utensils.	1. A table and chairs with wooden kitchen tools on top. 2. The table is full of wooden spoons and utensils. 3. A wood table holding an assortment of wood cooking utensils.
	1. A bathroom that has white towels in a rack over the tub. 2. Bathroom with a shower, sink, and toilet in it. 3. A bathroom with a sink, toilet and shower with curtain.	1. A very big white rest room with a shabby looking shower. 2. A bathroom that has white towels in a rack over the tub. 3. A bathroom with a toilet, towel rack and a tub in it.
A guy that is riding his bike next to a train.		
A man in a red shirt and a red hat is on a motorcycle on a hill side.		

Figure 6. Qualitative results on MS-COCO. For each query, we report top-3 ranked results. Predictions ordered by decreasing similarity score, with true matches are shown in blue. For text-to-image retrieval, the true and false matches are outlined in blue and red boxes, respectively.

Tasks	b_1 \ b_2	1.5	1.7	1.8	1.9
Video-to-Text	-0.2	36.0	35.5	36.1	35.5
	-0.3	34.6	35.0	**36.2**	35.6
	-0.4	34.0	35.3	35.6	35.3
Text-to-Video	-0.2	25.0	25.0	24.8	25.0
	-0.3	24.8	24.9	**25.0**	24.8
	-0.4	24.6	24.8	24.7	24.9

Table 5. The effect of b_1 and b_2 on MSR-VTT dataset.

COCO dataset. From the results, we find the average polynomial loss is converges faster than the max polynomial loss at the first few iterations. The reason is that the average polynomial loss contains more informative pairs. However, the final performance of max polynomial loss is slightly better than average polynomial loss. This is possibly due to the unreasonable parameter setting. Since average polynomial loss contains too many negatives pairs, it is difficult to find a set of parameters P_{Neg} to fit all informative negative pairs.

4.5. Qualitative Results

In this section, we perform visualizations Top-3 retrieval results for a handful of examples on MS-COCO. Both qualitative results of the image-to-text retrieval and the text-to-image retrieval are shown in Figure 6, which qualitatively illustrate the model behavior. Predictions are ordered by decreasing similarity score, with correct labels are shown in blue. From Figure 6, we can observe that by replacing the loss function with our polynomial loss, the performance of SCAN is further improved.

5. Conclusion

We have developed a universal weighting framework for cross-modal matching, which defines a weight function for the positive and negative pairs respectively. Universal weighting framework provides a powerful tool to analyze the interpretability of various loss functions. Furthermore, we proposed a polynomial loss function under the universal weighting framework, which can effectively sample and weight the informative pairs. Experimental results on four cross-modal matching benchmarks have demonstrated the proposed polynomial loss significantly improve the matching performance. In future work, we would like to investigate the potential of more advanced weighting functions for cross-modal matching.

Acknowledgments. This work was supported in part by the National Key Research and Development Program of China (No. 2018AAA0102200); the National Natural Science Foundation of China (No. 61976049 and No. 61632007); the Fundamental Research Funds for the Central Universities (No. ZYGX2019Z015); the Sichuan Science and Technology Program, China, (No. 2019ZDZX0008 and No. 2018GZDZX0032).

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[2] Hui Cui, Lei Zhu, Jingjing Li, Yang Yang, and Liqiang Nie. Scalable deep hashing for large-scale social image retrieval. *IEEE Transactions on Image Processing*, 29:1271–1284, 2019.

[3] Sounak Dey, Pau Riba, Anjan Dutta, Josep Llados, and Yi-Zhe Song. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*, pages 2179–2188, 2019.

[4] Jianfeng Dong, Xirong Li, and Cees GM Snoek. Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, 20(12):3377–3388, 2018.

[5] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, and Xun Wang. Dual dense encoding for zero-example video retrieval. *arXiv*, 2018.

[6] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*, 2018.

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[8] Chuang Gan, Tianbao Yang, and Boqing Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, pages 87–97, 2016.

[9] Chuang Gan, Yi Yang, Linchao Zhu, Deli Zhao, and Yueting Zhuang. Recognizing an action using its name: A knowledge-based approach. *International Journal of Computer Vision*, 120(1):61–77, 2016.

[10] Lianli Gao, Xiangpeng Li, Jingkuan Song, and Heng Tao Shen. Hierarchical lstms with adaptive attention for visual captioning. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[11] Jiuxiang Gu, Jianfei Cai, Shafiq R Joty, Li Niu, and Gang Wang. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *CVPR*, pages 7181–7189, 2018.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[13] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Similarity-based Pattern Recognition*, pages 84–92, 2015.

[14] Mengqiu Hu, Yang Yang, Fumin Shen, Ning Xie, Richang Hong, and Heng Tao Shen. Collective reconstructive embeddings for cross-modal hashing. *IEEE Transactions on Image Processing*, 28(6):2770–2784, 2018.

[15] Yan Huang, Qi Wu, Chunfeng Song, and Liang Wang. Learning semantic concepts and order for image and sentence matching. In *CVPR*, pages 6163–6171, 2018.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[17] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017.

[18] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018.

[19] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4654–4662, 2019.

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[21] Venice Erin Liong, Jiwen Lu, Yap-Peng Tan, and Jie Zhou. Deep coupled metric learning for cross-modal matching. *IEEE Transactions on Multimedia*, 19(6):1234–1244, 2016.

[22] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In *BMVC*, 2019.

[23] Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. Learning a recurrent residual fusion network for multimodal matching. In *ICCV*, pages 4107–4116, 2017.

[24] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *ACM MM*, pages 19–27, 2018.

[25] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. Dual attention networks for multimodal reasoning and matching. In *CVPR*, pages 299–307, 2017.

[26] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[27] Fumin Shen, Xiang Zhou, Jun Yu, Yang Yang, Li Liu, and Heng Tao Shen. Scalable zero-shot learning via binary visual-semantic embeddings. *IEEE Transactions on Image Processing*, 28(7):3662–3674, 2019.

[28] Heng Tao Shen, Luchen Liu, Yang Yang, Xing Xu, Zi Huang, Fumin Shen, and Richang Hong. Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[29] Jingkuan Song, Yuyu Guo, Lianli Gao, Xuelong Li, Alan Hanjalic, and Heng Tao Shen. From deterministic to generative: Multimodal stochastic rnns for video captioning. *IEEE transactions on neural networks and learning systems*, 30(10):3047–3058, 2018.

[30] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *CVPR*, pages 1979–1988, 2019.

[31] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *ACM MM*, pages 154–162, 2017.

[32] Tan Wang, Xing Xu, Yang Yang, Alan Hanjalic, Heng Tao Shen, and Jingkuan Song. Matching images and text with multi-modal tensor fusion and re-ranking. In *ACM MM*, pages 12–20, 2019.

[33] Jiwei Wei, Yang Yang, Jingjing Li, Lei Zhu, Lin Zuo, and Heng Tao Shen. Residual graph convolutional networks for zero-shot learning. In *Proceedings of the ACM Multimedia Asia*, pages 1–6. 2019.

[34] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016.

[35] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web*, 22(2):657–672, 2019.

[36] Xing Xu, Huimin Lu, Jingkuan Song, Yang Yang, Heng Tao Shen, and Xuelong Li. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. *IEEE transactions on cybernetics*, 2019.

[37] Xing Xu, Fumin Shen, Yang Yang, Heng Tao Shen, and Xuelong Li. Learning discriminative binary codes for large-scale cross-modal retrieval. *IEEE Transactions on Image Processing*, 26(5):2494–2507, 2017.

[38] Yang Yang, Jie Zhou, Jiangbo Ai, Yi Bin, Alan Hanjalic, Heng Tao Shen, and Yanli Ji. Video captioning by adversarial lstm. *IEEE Transactions on Image Processing*, 27(11):5600–5611, 2018.

[39] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[40] Bowen Zhang, Hexiang Hu, and Fei Sha. Cross-modal and hierarchical modeling of video and text. In *ECCV*, pages 374–390, 2018.

[41] Lei Zhang, Ji Liu, Yang Yang, Fuxiang Huang, Feiping Nie, and David Zhang. Optimal projection guided transfer hashing for image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.