Noisy Correspondence Learning with Self-Reinforcing Errors Mitigation

Zhuohang Dang1*, Minnan Luo1†, Chengyou Jia1, Guang Dai2,3, Xiaojun Chang4,6, Jingdong Wang5

1School of Computer Science and Technology, MOE KLINNS Laboratory, Xi’an Jiaotong University
2SGIT AI Lab
3State Grid Corporation of China
4University of Technology Sydney
5Baidu Inc
6Mohamed bin Zayed University of Artificial Intelligence

*This work was completed during his internship at SGIT AI Lab, State Grid Corporation of China.
†Corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Abstract

Cross-modal retrieval relies on well-matched large-scale datasets that are laborious in practice. Recently, to alleviate expensive data collection, co-occurring pairs from the Internet are automatically harvested for training. However, it inevitably includes mismatched pairs, i.e., noisy correspondences, undermining supervision reliability and degrading performance. Current methods leverage deep neural networks’ memorization effect to address noisy correspondences, which overconfidently focus on similarity-guided training with hard negatives and suffer from self-reinforcing errors. In light of above, we introduce a novel noisy correspondence learning framework, namely Self-Reinforcing Errors Mitigation (SREM). Specifically, by viewing sample matching as classification tasks within the batch, we generate classification logits for the given sample. Instead of a single similarity score, we refine sample filtration through energy uncertainty and estimate model’s sensitivity of selected clean samples using swapped classification entropy, in view of the overall prediction distribution. Additionally, we propose cross-modal biased complementary learning to leverage negative matches overlooked in hard-negative training, further improving model optimization stability and curbing self-reinforcing errors. Extensive experiments on challenging benchmarks affirm the efficacy and efficiency of SREM.

Introduction

Cross-modal matching, a key research area, focuses on retrieving relevant samples across various modalities. Contemporary methods achieve semantic alignment using modal-specific encoders (Diao et al. 2021; Li et al. 2021). They project data into a unified feature space, where matched data from different modalities are drawn together, while mismatched ones are pushed apart. To alleviate the laborious collection of well-matched data, recent datasets (Sharma et al. 2018) automatically collect co-occurring sample pairs from the Internet for training. However, they contain around 20% mismatched pairs (Sharma et al. 2018; Huang et al. 2021), namely noisy correspondences. Encouraging these mismatched pairs to be similar will significantly degrade the matching performance.

Figure 1: Drawback illustrations of previous methods.

Recent advancements (Yang et al. 2023) have tackled noisy correspondences through neural network’s memorization, which enables clean samples to exhibit higher similarities than noisy ones after the initial few epochs (Yao et al. 2020). Specifically, after warmup, these methods further re-
fine similarity prediction with the following alternate steps: 1) Using similarity scores to identify clean samples. 2) Deriving soft margins proportional to similarity scores for robust matching of selected clean samples. The soft margins are employed in a hinge-based ranking loss, where a larger margin intensifies the model’s sensitivity towards differentiating the given sample from its negatives. However, Figure 1(a) shows that such an approach is susceptible to self-reinforcing errors. The primary vulnerability arises from that clean sample selection and corresponding sensitivity estimation rely heavily on model’s similarity prediction. This leads to a critical issue where confident but incorrect similarity predictions are amplified during subsequent training, forming a loop of self-reinforcing errors (Chen et al. 2023; Yang et al. 2023). Furthermore, hinge-based ranking loss solely focuses on query’s positive and hard negative sample, overlooking numerous negative information. Figure 1(b) shows that this narrow focus results in suboptimal model optimization, potentially aggravating self-reinforcing errors.

In light of above, we propose a novel noisy correspondence learning framework, namely SREM, with three core modules: 1) We introduce a novel energy-guided approach to complement conventional similarity-based sample filtration. We first produce classification logits for a sample by viewing sample matching as a classification task within the batch. We then use energy scores derived from classification logits to gauge the model’s uncertainty during sample selection. As a result, this strategy ensures the selected clean samples maintain both high similarity and low uncertainty, paving the way for more precise data division. 2) We propose a Swapped Gradient Weighting (SGW) strategy. SGW assesses the model’s sensitivity towards individual samples by leveraging swapped classification entropy for robust matching. Samples with lower entropy suggest higher prediction confidence, thus the model should be more sensitive to them and let them contribute more to optimization (Iscen et al. 2019). In contrast to a single similarity score, SGW considers the model’s prediction distribution over both clean and negative samples, ensuring robustness. 3) We introduce a novel Cross-Modal Biased Complementary Learning (CMBCL) objective for leveraging negative samples overlooked in the hinge-based ranking loss. We perceive these overlooked negative matches as “complementary labels” that essentially signal non-matching samples, guiding the model to distance positive samples from all negatives and thus circumventing potential self-reinforcing errors.

Extensive experiments highlight that SREM surpasses state-of-the-art by more than 1% in average recall and reduces training time by more than 40%. Moreover, we theoretically prove CMBCL’s efficacy, as it converges to an optimal classifier equivalent to one trained with true labels. We also highlight its generality by encompassing the strong competitor RCL (Hu et al. 2023) as a special case.

**Related Work**

**Cross Modal Retrieval**

Cross-modal matching (Radford et al. 2021; Chen et al. 2021; Diao et al. 2021; Lee et al. 2018) aims to project images and texts into a unified space where matched multi-modal pairs are similar while mismatched are dissimilar.

Contrary to previous approaches that presuppose well-matched training data, the prohibitive collection costs have fostered the emergence of new paradigms like noisy correspondences, a prevalent issue in domains such as person re-id (Yang et al. 2022a), graph matching (Lin et al. 2023), and multi-view learning (Yang et al. 2022b, 2021). Current methods in cross-modal matching (Yang et al. 2023; Han et al. 2023; Huang et al. 2021) primarily employ multi-step frameworks: They first estimate the distribution of instance-level loss/similarity across the entire dataset. Then they compute the posterior probability as the pseudo-label for each sample, which is further filtered by a threshold and clean samples are used for training. To eliminate additional computation overhead caused by similarity distribution estimation, DECL (Qin et al. 2022) uses similarity with evidential learning to dynamically filter out noisy correspondences within each batch. However, similarity-guided training in previous methods lead to self-reinforcing errors. In contrast, our SREM addresses overconfidence in similarity scores through overall prediction distributions, effectively mitigating such errors and notably enhancing performance.

**Complementary Label Learning**

Unlike conventional classification tasks, samples in complementary label learning (CLL) are assigned complementary labels that indicate classes they do not belong to. To effectively use these weak supervisions, (Ishida et al. 2017, 2019) assume the uniform distribution of complementary labels and prove an optimal classifier can be learned with mere complementary labels. Differently, some works (Yu et al. 2018; Gao and Zhang 2021; Xu et al. 2020) consider the unknown distribution of complementary labels. By estimating label transition probabilities, they inferred the distribution of complementary labels and subsequently refined them for training. In noisy correspondence learning, RCL (Hu et al. 2023) extends CLL to introduce a novel contrastive learning framework that exclusively leverages negative information, mitigating the potential negative effects of mismatched samples. However, the neglect of powerful positive supervision leads to suboptimal results for RCL. On the contrary, beyond using positive supervision in ranking loss, we additionally leverage the dissimilarity of negative samples to utilize negative information more effectively, therefore achieving a more robust training regime against noisy correspondence.

**Methodology**

**Problem Definition**

Following previous works, we use image-text retrieval as a proxy task to explore noisy correspondence in cross-modal matching, consisting of two sub-tasks: image-to-text (t2i) and text-to-image (t2i) retrieval. Let \( D = \{(I_i, T_i, m_i)\}_{i=1}^N \) denote a training dataset, where \( N \) is the data size and \((I_i, T_i)\) is the \( i \)-th image-text pair with label \( m_i \in \{0, 1\} \) indicating whether they are matched. In noisy correspondence, an unknown portion of pairs in \( D \) is mismatched, i.e., the image and text are not matched but with matched labels.
Model Overview

In this section, we detailly present our SREM with an overview shown in Figure 2. For simplicity, we take image-to-text retrieval as a showcase to introduce the pipeline of SREM, while text-to-image retrieval is conducted symmetrically. Initially, the feature encoder generates similarity logits from the input pair. Then, we employ three elaborately-designed modules to mitigate the self-reinforcing errors during training. Given the disparities in prediction distribution, we utilize energy uncertainty to segregate clean samples, denoted as \( \mathcal{D}_{\text{clean}} \), from noisy correspondences, \( \mathcal{D}_{\text{noisy}} \). To enhance SREM’s robustness, we introduce the swapped gradient weighting and cross-modal biased complementary learning framework. The former proposes a gradient-rescaled ranking loss \( L_u \), while the latter effectively leverages the overlooked negative matches \( \mathcal{D}_{\text{neg}} \) in \( L_w \) as complementary labels. We will detail each component and corresponding optimization objective in what follows.

Feature Encoder

Initially, the feature encoder projects both visual and textual data into a unified feature space using model-specific encoders \( f \) and \( g \), respectively. Within the unified feature space, a function \( h \) computes the similarity logit as \( F_{ij} = h(f(I_i), g(T_j)) \) (\( h(I_i, T_j) \) for short), where the corresponding similarity score is defined as \( S_{ij} = \sigma(F_{ij}) \). Here, \( \sigma(\cdot) \) denotes the sigmoid activation function.

Energy-Guided Sample Filtration

Our objective is to circumvent the pitfalls of previous methods that overconfidently divide samples with similarity prediction, thereby introducing potential sample selection risk. Take the similarity scores \([0.85, 0.80, 0.82]\) as an example: the first score represents the given sample pair, while the others correspond to its negative samples. Even though the given sample pair exhibits a high similarity score, it is not significantly different from the negative samples, suggesting a possible mismatch. Hence, selecting such a sample pair as “clean” based solely on similarity can be risky. To this issue, by considering the overall prediction distribution, we aim to explore sample selection uncertainty to complement similarity-based sample filtration. Given a batchsize \( B \), we first generate the classification logits \( F_i \) of the visual input \( I_i \) by viewing sample matching as a classification task within the batch. \( F_i \) is formulated as \( F_i = \{F_{i1}, \ldots, F_{iB}\} \) with a corresponding label \( y_i = i \). Due to DNN’s memorization effect, the model initially becomes adept at recognizing clean samples, leading to an unimodal distribution at \( y_i \). In contrast, model struggles to differentiate noisy correspondences from their negatives, also giving rise to a more uniform distribution.

In view of such difference, we turn to energy uncertainty in logits space, which is a widely acceptable metric in the literature of uncertainty learning (Liu et al. 2020; Xie et al. 2022). Specifically, the energy uncertainty corresponding to the visual input \( I_i \) can be calculated by:

\[
\text{Energy}(I_i) = -\log \sum_{b=1}^{B} e^{F_{ib}}.
\]  

Intuitively, more uniformly distributed prediction \( i.e., \) noisy correspondence \( \) leads to higher estimated energy uncertainty (Zhang et al. 2023). Therefore we select the clean samples by applying a threshold \( \tau \) and the maximum similarity constraint (Qin et al. 2022), \( i.e., \)

\[
\mathcal{D}_{\text{clean}} = \{ i | \text{Energy}(I_i) < \tau \text{ and } y_i = \arg \max_j F_{ij} \}, \quad (2)
\]

while \( \mathcal{D}_{\text{noisy}} \) refers to mismatched samples. In this sense, the selected samples maintain both low uncertainty and high similarity, paving the way for more precise sample division.

Moreover, we conceive an energy-bounded loss \( L_u^i \) to reduce the energy uncertainty of clean samples while enhancing that of noisy samples for enlarged margins, \( i.e., \)

\[
L_u^i = \mathbb{E}_{I_i \sim \mathcal{D}_{\text{clean}}} \left[ 0, \text{Energy}(I_i) - m_{\text{clean}} \right]_+^2 + \mathbb{E}_{I_i \sim \mathcal{D}_{\text{noisy}}} \left[ 0, m_{\text{noisy}} - \text{Energy}(I_i) \right]_+^2, \quad (3)
\]

where \( [x]_+ = \max(x, 0) \); \( m_{\text{clean}} \) and \( m_{\text{noisy}} \) are separate margins that penalize the clean (noisy) samples with energy uncertainty higher (lower) than the given margin.

Swapped Gradient Weighting

After sample filtration, it is risky to directly train the model on \( \mathcal{D}_{\text{clean}} \) as it potentially contains some false positives (Huang et al. 2021; Yang et al. 2023). To ensure robust training, it’s crucial to devise strategies that allow the model to adaptively maintain varied sensitivities to samples within \( \mathcal{D}_{\text{clean}} \). Instead of overconfident single similarity score, we introduce classification entropy to estimate sensitivity of each clean sample. Visual input \( I_i \)’s classification distribution is defined as \( P_i = \text{softmax}(F_i) \), and the corresponding normalized classification entropy \( e(P_i) \) is formulated as:

\[
e(P_i) = -\frac{\sum_{j=1}^{B} P_{ij} \log P_{ij}}{\log B}.
\]  

Here \( \log B \) is the maximum entropy to scale \( e(P_i) \) into \([0, 1]\) for numerical stability. In this sense, low \( e(P_i) \) highlights the model’s ability to recognize matched samples, while suppressing similarity scores to other negative samples. Consequently, model should be more sensitive to samples with lower \( e(P_i) \) in optimization (Iscen et al. 2019).
In light of above, let \( w^T_t \) denote the entropy-based model’s sensitivity to visual input \( I_t \) in i2t retrieval, formulated by:

\[
w^T_t = 1 - e(P_t) \mathbb{I}(\alpha - S_i + \sigma(h(I_t, T_{\phi(i)}))],
\]

where \( \alpha > 0 \) is the expected margin between positive and negative match; \( \phi(i) = \arg \max_{j \neq i} F_{ij} \) and \( T_{\phi(i)} \) is the hard negative text of \( I_t \), i.e., the negative text most similar to \( I_t \) within the batch. Moreover, we employ indicator function \( \mathbb{I}() \) to evaluate whether a sample and its hard negative have expected discrimination \( \alpha \). This design avoids unnecessary gradients on samples exhibiting satisfactory discrimination, reducing the risk of overfitting. Besides, we further employ swapped prediction strategy on calculated \( e(P_t) \), which is widely used in cross-modal tasks for improving robustness (Andonian, Chen, and Hamid 2022). Its key idea is to use the weights derived from one modality for the other modality, promoting cross-modal consistency in the learning process. For example, we use \( w^T_t \) derived from t2i classification entropy for i2t retrieval training, and vice-versa. Specifically, we apply \( w^T_t \) with hinge-based ranking loss, defined as:

\[
L_w^{T2t} = \mathbb{E}_{I_t \sim D_{clean}}[\alpha - w^T_t S_{ii} + \sigma(h(I_t, T_{\phi(i)}))] + \frac{\sigma}{\alpha},
\]

As a result, the derivative of \( L_w^{T2t} \) with respect to model parameters \( \theta \) is given by the chain rule \( \frac{\partial L_w^{T2t}}{\partial \theta} = \frac{\partial L_w^{T2t}}{\partial S_{ij}} \frac{\partial S_{ij}}{\partial \theta} \) with

\[
\frac{\partial L_w^{T2t}}{\partial S_{ij}} = \begin{cases} 
  w^T_t, & j = i \\
  -1, & j = \phi(i) \\
  0, & \text{otherwise}
\end{cases}.
\]

Equation (7) implies that clean samples exhibiting more certain distributions will retain larger gradients, consequently to which model is more sensitive. Compared to sample-reweighting methods (Wei et al. 2021; Wang et al. 2019), our SGW strategy further suppresses similarity scores to hard negatives as \( \frac{\partial L_w^{T2t}}{\partial S_{\phi(i)}} = -1 \). Thus, Equation (6) can effectively adjust model’s sensitivity of different samples in optimization, enhancing matching robustness.

### Cross-Modal Biased Complementary Learning

Evidently, Equation (7) highlights that \( L_w^{T2t} \) overlooks numerous negative similarities defined as:

\[
D_{neg} = \{S_{ij} \mid j \neq i; \text{ and if } i \in D_{clean}, j \neq \phi(i)\}. \tag{8}
\]

These overlooked negative similarities maintain zero gradients and are ignored in model optimization. However, in classification, these overlooked similarities indicate the samples that do not match the given sample, i.e., complementary labels. As shown in Figure 2, harnessing these complementary labels can enhance the stability of the model optimization. In this sense, we construct an auxiliary dataset \( D_{neg} = \{(i, \bar{y}_i)\}_{i=1}^{B} \) within each batch. Here \( i \) is the index of given image \( I_t \) within batch, \( \bar{y}_i \) is corresponding complementary labels formulated as:

\[
\bar{y}_i = \{0, 1, \cdots, B - 1\} \setminus \{y_i\}. \tag{9}
\]

Furthermore, we explore non-uniformly distributed complementary labels to improve model’s generality, due to the following facts: 1) Ideal uniformly distributed complementary labels do not necessarily hold in real-world data, particularly in instance-level classification. 2) Non-uniform complementary labels permit model to focus more on the harder negatives, thereby preventing informative supervision from being overwhelmed by redundant negative samples.

Inspired by (Yu et al. 2018; Gao and Zhang 2021), we prefer to choose negative with higher similarity as the complementary label, enabling model to focus more on challenging and informative negative counterparts. Notably, we directly use similarity to estimate the selection probability of complementary labels, as they are leveraged to suppress negative information and do not involve self-reinforcing errors. Specifically, we employ MS (Wang et al. 2019) to gauge the likelihood of selecting \( T_j \) as \( I_t \)’s complementary label, considering both self and relative similarities as:

\[
\bar{p}_{ij}^{\phi} = \frac{e^{\beta(S_{ij}-b)}}{1 + \sum_{k \in \bar{S}_i} e^{\beta(S_{ik}-b)}}, \tag{10}
\]

where \( \beta \) and \( b \) are two hyperparameters of Binomial deviance (Hastie et al. 2009), controlling the smoothness of selection distribution. Note that selected hard negatives from \( D_{clean} \) have already been considered in \( L_w^{T2t} \), we exclude these samples to prevent their over-representation in the model training process, which is formulated by:

\[
\bar{p}_{\phi(i)}^{ij} = -\infty, \forall i \in D_{clean}. \tag{11}
\]

We then rectify complementary labels using the overall selection probability (Yu et al. 2018), i.e.,

\[
S' = \text{softmax}(\bar{p}^{T2t})^T S. \tag{12}
\]

Ultimately, the cross-modal biased complementary learning objective \( L_c^{T2t} \) on \( D_{neg} \) is formulated as:

\[
L_c^{T2t} = -\mathbb{E}_{(i, \bar{y}_i) \sim D_{neg}} \mathbb{E}_{j \sim \bar{y}_i} \log(1 - S'_{ij}). \tag{13}
\]

Moreover, we provide theoretical evidence to better elucidate CMBCL’s efficacy in Theorem 1.

**Theorem 1.** Given sufficient data with complementary labels, minimizing Equation (13) can yield the optimal classifier equivalent to that trained with the true labels.

### Model Optimization

To ensure consistent performance across modalities, we employ SREM for bidirectional matching, encompassing both image-to-text and text-to-image tasks, formulated by:

\[
\min_{\theta} L = 0.5(L_u^{T2t} + L_t^{T2t}) + \lambda_1(L_u^T + L_t^T) + \lambda_2(L_c^T + L_c^{T2t}),
\]

where \( L_u^T, L_t^{T2t}, \) and \( L_c^{T2t} \) represent objectives when symmetrically applying energy-guided sample filtration, SGW, and CMBCL for text-to-image retrieval. \( \lambda_1, \lambda_2 \in [0, 1] \) are hyperparameters to adjust the effect of energy uncertainty estimation and negative information utilization.

### Experiments

#### Experiments Setting

**Datasets** Following previous works (Han et al. 2023), we evaluate SREM using three image-text retrieval datasets, including COCO (Lin et al. 2014), Flickr30K (Young et al. 2015), and MS-COCO.
Table 1: Image-Text Retrieval on Flickr30K and MS-COCO 1K. Results marked with ‘*’ are reproduced results from their official code, while ‘†’ signifies methods that incorporate additional priors.

<table>
<thead>
<tr>
<th>Noise</th>
<th>Methods</th>
<th>Flickr30K</th>
<th>MS-COCO</th>
<th>Flickr30K</th>
<th>MS-COCO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Image→Text</td>
<td>Text→Image</td>
<td>Image→Text</td>
<td>Text→Image</td>
<td></td>
</tr>
<tr>
<td>20%</td>
<td>SCAN</td>
<td>58.5</td>
<td>2.8</td>
<td>84.7</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td>VSRN</td>
<td>33.4</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>IMRAM</td>
<td>22.7</td>
<td>14.6</td>
<td>48.2</td>
<td>14.6</td>
</tr>
<tr>
<td></td>
<td>SAF</td>
<td>62.8</td>
<td>4.2</td>
<td>98.6</td>
<td>4.2</td>
</tr>
<tr>
<td></td>
<td>SGR</td>
<td>55.9</td>
<td>7.6</td>
<td>98.2</td>
<td>7.6</td>
</tr>
<tr>
<td></td>
<td>NCR</td>
<td>73.5</td>
<td>11.9</td>
<td>98.2</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>DECL</td>
<td>77.5</td>
<td>11.9</td>
<td>98.2</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>BiCro†</td>
<td>78.1</td>
<td>11.9</td>
<td>98.2</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>MSCN*†</td>
<td>77.4</td>
<td>11.9</td>
<td>98.2</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>RCL</td>
<td>75.9</td>
<td>11.9</td>
<td>98.2</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>79.5</td>
<td>11.9</td>
<td>98.2</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>SCAN</td>
<td>26.0</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>VSRN</td>
<td>2.6</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>IMRAM</td>
<td>5.3</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>SAF</td>
<td>7.4</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>SGR</td>
<td>4.1</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>NCR</td>
<td>68.1</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>DECL</td>
<td>72.7</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>BiCro†</td>
<td>74.6</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>MSCN*†</td>
<td>71.9</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>RCL</td>
<td>72.7</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>76.5</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td>40%</td>
<td>SCAN</td>
<td>13.6</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>VSRN</td>
<td>0.8</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>IMRAM</td>
<td>1.5</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>SAF</td>
<td>0.1</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>SGR</td>
<td>1.5</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>NCR</td>
<td>13.9</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>DECL</td>
<td>65.2</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>BiCro†</td>
<td>67.6</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>MSCN*†</td>
<td>67.5</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>RCL</td>
<td>67.7</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
<tr>
<td></td>
<td>Ours</td>
<td>71.0</td>
<td>11.9</td>
<td>51.4</td>
<td>11.9</td>
</tr>
</tbody>
</table>

Comparison with State-Of-The-Art

We compare the proposed SREM against current state-of-the-art (SOTA) methods to demonstrate its effectiveness, including general image-text retrieval methods SCAN (Lee et al. 2018), VSRN (Li et al. 2019), IMRAM (Chen et al. 2020), SGR, SAF (Diao et al. 2021), and noisy correspondence robust methods NCR (Huang et al. 2021), DECL (Qin et al. 2022), MSCN (Han et al. 2023), BiCro (Yang et al. 2023) and RCL (Hu et al. 2023).

Results on Synthetic Noise of Flickr30K and MS-COCO

As in previous works, we emulate noisy correspondences by randomly shuffling the training images and captions for

SGRAF (Diao et al. 2021), with the same training settings as (Huang et al. 2021) for fair comparisons. Specifically, we warm up the model for 5 epochs with $L_2$ and $L_2$ to achieve initial convergence, followed by a 50 epochs training process. We employ a batch size of 128 and anAdam (Kingma and Ba 2014) optimizer with a learning rate of $2e-4$ that will be decayed by 0.1 after 25 epochs.
Figure 3: We visualize the energy uncertainty distribution of clean and noisy pairs at different training stages of our SREM, which is conducted on Flickr30K under 20% noise. Thanks to SREM, the energy uncertainty of clean pairs gradually approaches the left (low) and the energy uncertainty of noisy pairs tightly gathers to the right (high).

Table 2: Image-Text Retrieval on CC152K.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Image→Text</th>
<th>Text→Image</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSRN</td>
<td>32.6/61.3/70.5</td>
<td>32.5/59.4/70.4</td>
<td>326.7</td>
</tr>
<tr>
<td>IMRAM</td>
<td>33.1/57.6/68.1</td>
<td>29.0/56.8/67.4</td>
<td>312.0</td>
</tr>
<tr>
<td>SAF</td>
<td>31.7/59.3/68.2</td>
<td>31.9/59.0/67.9</td>
<td>318.0</td>
</tr>
<tr>
<td>NCR</td>
<td>39.5/64.5/73.5</td>
<td>40.3/64.6/73.2</td>
<td>355.6</td>
</tr>
<tr>
<td>DECL</td>
<td>39.0/66.1/75.5</td>
<td>40.7/66.3/76.7</td>
<td>364.3</td>
</tr>
<tr>
<td>BiCro</td>
<td>40.8/67.2/76.1</td>
<td>42.1/67.6/76.4</td>
<td>370.2</td>
</tr>
<tr>
<td>MSCN</td>
<td>40.1/65.7/76.6</td>
<td>40.6/67.4/76.3</td>
<td>366.7</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>40.9/67.5/77.1</strong></td>
<td><strong>41.5/68.2/77.0</strong></td>
<td><strong>372.2</strong></td>
</tr>
</tbody>
</table>

Table 3: Component Analyses on CC152K with real-world noise. \( L_u^c \) denotes conventional complementary learning without considering the complementary label distribution.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Image→Text</th>
<th>Text→Image</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( L_u )</td>
<td>( L_w )</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
</tr>
<tr>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
<td>✓ ✓ ✓</td>
</tr>
</tbody>
</table>

Table 3: Component Analyses on CC152K with real-world noise. \( L_u^c \) denotes conventional complementary learning without considering the complementary label distribution.

Ablation Studies

Component Analyses

Table 3 shows that vanilla trained model exhibits suboptimal performance, illustrating its susceptibility to disturbances caused by noisy correspondences. The energy-guided sample filtration significantly enhances the performance by more than 10% on \( R\% \). When using swapped gradient weighting, we observe performance boosts in all results, except \( R\% \) for text retrieving. Furthermore, with consideration of label distributions, leveraging unused negative information as biased complementary labels considerably improves performance, evidenced by an increase of more than 1% in \( R\% \). These results underline the significant role of complementary labels in fortifying retrieval robustness. The best performance is achieved with all proposed components, demonstrating their efficacy.

Visualization on Energy Uncertainty

Figure 3 visualizes energy uncertainty during training. As training progresses, the energy uncertainty of clean samples becomes lower while that of noisy correspondences increases, manifesting a clear polarizing trend. These observations validate the efficacy of energy uncertainty estimation for noisy correspondences. Therefore, the energy uncertainty from the overall prediction distribution can naturally be used to differentiate between noisy and clean pairs, further boosting the robustness against noisy correspondences.

except for \( R\% \) of retrieving images. These results demonstrate SREM’s appealing efficacy in real-world scenarios.
Visualization on Self-Reinforcing Errors  We track the training progress to validate the efficacy of our SREM in alleviating self-reinforcing errors. Specifically, we measure the performance of each epoch, as well as noisy gradients ratio relative to all positive gradients (the proportion of false positive gradients created by enhancing mismatched samples’ similarity in \( L_w \)). We also provide the results of MSCN and BiCro for more comprehensive and fair comparisons. As shown in Figure 5, since CMBCL during warmup avoids self-reinforcing errors, SREM starts with the lowest noisy gradient ratio. While in training, with its carefully designed components, SREM effectively suppresses self-reinforcing errors, exhibiting a significantly lower and stable noise gradient ratio, i.e., less than 7%. In contrast, MSCN and BiCro start with higher noise gradient ratios that rapidly increase in training due to their similarity-based training with hard negatives. As a result, SREM achieves better performance with stable optimization, while MSCN and BiCro exhibit unsatisfactory results, whose performance gradually drops with noisy gradient ratio increasing. These results highlight the efficacy of SREM in alleviating self-reinforcing errors.

Efficiency Analyses  We report the training overhead per epoch on CC152K using an NVIDIA Tesla A40 48G in Table 4. The training time of MSCN and BiCro contains two parts as they first pre-compute similarity across the entire dataset and then conduct sample filtration before training. These steps incur additional computation and storage overhead. Moreover, MSCN computes meta gradients for model optimization and BiCro rectifies soft correspondences via numerous anchor samples, both of which are computationally expensive and thus further diminishing efficiency. Differently, our SREM not only eliminates the pre-computation but also employs computationally efficient techniques, i.e., energy uncertainty, entropy and complementary learning. Consequently, SREM reduces the training time by more than 40%, highlighting its efficiency and potential applicability to large-scale datasets.

Detected Noisy Real-World Correspondences  Figure 4 shows some real-world noisy correspondences in CC152K detected by our SREM with their corresponding energy uncertainty. Specifically, SREM is not limited to recognizing only obvious noisy pairs containing completely irrelevant information. It also can identify hard mismatched pairs with subtle semantic misalignment, e.g., the missing elements of the phone, hands, steak and fountain, etc., as well as the incongruence between concepts like “building” and “industry”. These results qualitatively demonstrate SREM’s efficacy for handling real-world applications.

Conclusion

This paper presents a novel framework, SREM, to address the challenges of noisy correspondences in cross-modal matching. Using per-sample classification logits, SREM ingeniously employs energy uncertainty to filter out the noisy correspondences, paving the way for more precise data division. It then applies SGW to recalibrate gradients, offering a more nuanced approach to assessing model’s sensitivity in sample matching. Moreover, the CMBCL framework within SREM harnesses previously overlooked negative information, ensuring stable model optimization. Both theoretical evidence and extensive experiments on challenging benchmarks corroborate SREM’s superiority in efficacy, efficiency and generality. We hope our SREM will drive improvements in both the efficacy and efficiency of noisy correspondence learning, providing new insights into building more robust cross-modal information retrieval systems.
Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2022YFB3102600), National Nature Science Foundation of China (No. 62192781, No. 62272374, No. 62202367, No. 62250009, No. 62137002), Project of China Knowledge Center for Engineering Science and Technology, Project of Chinese academy of engineering “The Online and Offline Mixed Educational Service System for ‘The Belt and Road’ Training in MOOC China”, and the K. C. Wong Education Foundation.

References


