Differential-informed Sample Selection Accelerates Multimodal Contrastive Learning

Supplementary Material

Contents

A. Broader Impact

This paper focuses on accelerating multimodal contrastive learning through a sample selection strategy, an approach that is crucial during the pre-training phase. This challenge is particularly prevalent when training on large-scale, real-world multimodal datasets, where time and resource constraints are often limiting factors. Accelerating contrastive learning is not only valuable for vision-language pre-training, but also has far-reaching implications for a variety of multimodal applications, including video, audio, and medical fields.

In this paper, we propose accelerating the learning process by preventing learning from noisy correspondence samples through an oracle-free, online batch selection method. While several existing approaches aim at addressing noisy correspondence, they primarily rely on dual-network structures, which incur significant computational costs. In contrast, our proposed method, DISSect, effectively eliminates noisy correspondences while accelerating the learning procedure. DISSect is both simple to implement and highly effective, and this research direction holds great promise for further enhancement through more detailed and refined studies.

B. Limitation and Future Explorations

While the proposed DISSect method shows promising results, there are several limitations that warrant attention. One key limitation is that DISSect has been implemented and evaluated solely in dual-modality scenarios. However, many real-world multimodal scenarios involve more than two modalities, such as combinations of video, audio, images, and text, which introduce higher computational costs and exacerbate the issue of noisy correspondence. Moreover, DISSect can be further improved. For example, previous research has shown that hard samples can accelerate multimodal model training during the earlier stages, which is a consideration not explored in this paper. Besides, the extra forward propagation required by online batch selection paradigm can also be accelerated through further low-level optimizations. Therefore, we call for future research to address these challenges in more depth and develop more refined strategies to further accelerate multimodal contrastive learning.

C. Details of Theoretical Insights

In this section, we give demonstration to the theoretical insight in the main paper. The theoretical insight is based on the theorem of the memorization effect in [40]. Here we notate clean data pairs as $\{I_i, T_i\}$ and noisy data pairs as $\{\tilde{I}_i, \tilde{T}_i\}$. Specifically, during the early stage of training, the gradient descent of loss $-\nabla\mathcal{L}$ points approximately in the average direction of all samples. Since the majority of data samples are correctly corresponded, such gradient is well correlated with the correct direction during this stage.

Once at the early learning point e when most clean patterns have been learned by the model, the gradient of clean samples $\frac{\partial\mathcal{L}}{\partial f(I_i)^\top g(T_i)}$ approaches zero. Meanwhile, the noisy correspondence samples have not been learned, with gradient of noisy samples $\frac{\partial\mathcal{L}}{\partial f(\tilde{I}_i)^\top g(\tilde{T}_i)}$ remains large. The gradient begins to point in directions orthogonal to the correct direction of clean samples. When the dimension of parameter space is sufficiently large, there are enough of these orthogonal directions to allow the model to completely memorize the noisy patterns, until gradient of all samples approach zero. As a result, we have the following conclusion,

◦ For $t < e$, $-\nabla\mathcal{L}(\theta_1^t, \theta_2^t)$ is well correlated with the correct direction of clean samples, and at $t = e$ the model learns a better similarity between clean samples than noisy samples.

◦ At $t = e$, $\frac{\partial\mathcal{L}}{\partial f(I_i)^\top g(T_i)}$ vanishes to approaching zero, while $\frac{\partial\mathcal{L}}{\partial f(\tilde{I}_i)^\top g(\tilde{T}_i)}$ remains a large value.

◦ As $t > e \to T$, the model memorizes all samples as $\frac{\partial\mathcal{L}}{\partial f(I_i)^\top g(T_i)}$ for all samples approach zero.

In the following part, we demonstrate that the differential between CLIPScore inherently reflects the gradient changing tendency in $t \in [e, T]$. Firstly, we calculate the partial of loss $\mathcal{L}$ to similarity score $f(I_i)^\top g(T_i)$.

$$\mathcal{L}_{f\to g} = -\frac{1}{|\mathcal{D}_b|}\sum_{i=1}^{|\mathcal{D}_b|}\log\frac{\exp\left(f(I_i)^\top g(T_i)/\tau\right)}{\sum_{j=1}^{|\mathcal{D}_b|}\exp\left(f(I_i)^\top g(T_i)/\tau\right)}, \tag{6}$$

$$\frac{\partial\mathcal{L}_{f\to g}}{\partial f(I_i)^\top g(T_i)} = \frac{1}{|\mathcal{D}_b|}\sum_{i=1}^{|\mathcal{D}_b|}(1 - \frac{\exp\left(f(I_i)^\top g(T_i)/\tau\right)}{\sum_{j=1}^{|\mathcal{D}_b|}\exp\left(f(I_i)^\top g(T_i)/\tau\right)}) \tag{7}$$

Since the positive pairs have been well-learned to contrast to negative pairs in the early stage, CLIPScore that refers to $\text{CLIPScore}(I_i, T_i) = w * \max\left(f(I_i)^\top g(T_i), 0\right)$ maintains a positive value as $f(I_i)^\top g(T_i)$ during this stage. Here we

further contend that the similarity sum of negative pairs maintains unchanged during $t \in [e, T]$ and can be regard as a constant $\sum_{j=1,j\neq i}^{|\mathcal{D}_b|} \exp\left(f(I_i)^\top g(T_i)/\tau\right) \approx C$. Then we can infer that the partial gradient of contrastive loss to $f(I_i)^\top g(T_i)$ is negatively correlated with the exponential of CLIPScore.

$$\frac{\partial \mathcal{L}_{f\to g}}{\partial f(I_i)^\top g(T_i)} \approx \frac{1}{|\mathcal{D}_b|}\sum_{i=1}^{|\mathcal{D}_b|}\frac{C}{\exp\left(f(I_i)^\top g(T_i)/\tau\right)+C} \propto \frac{1}{|\mathcal{D}_b|}\sum_{i=1}^{|\mathcal{D}_b|}\exp\left(\text{CLIPScore}\right)^{-1} \tag{8}$$

The differential of gradient between $t = e$ and $t = T$ indicates the changing tendency of gradient during this stage, which equals as the following function,

$$\begin{aligned}\frac{\partial \mathcal{L}_{f\to g,[e]}}{\partial f(I_i)^\top g(T_i)} - \frac{\partial \mathcal{L}_{f\to g,[T]}}{\partial f(I_i)^\top g(T_i)} &\approx \frac{1}{|\mathcal{D}_b|}\sum_{i=1}^{|\mathcal{D}_b|}\Big(\frac{C}{\exp\left(f(I_i)^\top g(T_i)/\tau\right)+C}[e] - \frac{C}{\exp\left(f(I_i)^\top g(T_i)/\tau\right)+C}[T]\Big) \\ &= \frac{1}{|\mathcal{D}_b|}\sum_{i=1}^{|\mathcal{D}_b|} k\cdot\left(\exp(\text{CLIPScore}[e])^{-1} - \exp(\text{CLIPScore}[T])^{-1}\right)\end{aligned} \tag{9}$$

where k is the negative correlation coefficient. Such deduction explains that the negative differential of CLIPScore reflects the changing tendency of gradient. During $t \in [e, T]$, for clean samples, the gradient maintains approximately zero since the clean patterns have been learned until $t = e$, leading to a relative small discrepancy in gradient change. For noisy samples, since such noisy patterns have not been memorized at $t = e$, the gradient remains a large value at the early learning point up until $t = T$ when noisy patterns are wrongly memorized with gradient approaching zero. As a result, the discrepancy of gradient in the equation evidently increases during the late training, which is clearly contrasted to the clean samples. DISSect successfully identifies the noisy correspondence samples by discriminating on the differential of CLIPScore that captures this contrast in gradient changing tendency.

D. Other Related Work

In this section, we discuss about other efficient training strategies besides sample selection. Since they are less relevant to our paper, we put this part in the appendix.

D.1. Curriculum Learning

Curriculum learning has been widely studied to improve model performance by progressively increasing task complexity. In visual question answering, Akl et al. [2] breaks down the VQA task into smaller sub-tasks based on question types and trains the model on a sequence of progressively harder tasks. The C-SFDA [27] framework for source-free domain adaptation employs curriculum learning to select pseudo-labels based on their reliability, improving the adaptation process. DoCL [72] optimizes learning dynamics by focusing on samples at the learning frontier—those with large loss but high learning potential. Zhou and Bilmes [71] adaptively selects subsets of training data at various stages to provide a balance of task difficulty. Uncertainty-aware curriculum learning [73] in neural machine translation adjusts curriculum based on the model's uncertainty, using cross-entropy and weight variance to determine data difficulty. In conclusion, sample selection has learned from the idea of curriculum learning and selects samples in a more adaptive manner.

D.2. Dataset Distillation

Dataset distillation optimizes training data for more efficient neural network training. Anil et al. [4] accelerates training and improves model accuracy in large-scale distributed networks through online distillation. Li et al. [33] uses contrastive loss to align image and text representations before fusing them, enhancing vision-language learning. Radenovic et al. [50] applies filtering and concept distillation to leverage unimodal representations for contrastive training. Cazenavette et al. [6] optimizes synthetic data by minimizing the distance between parameters trained on synthetic and real data. Cui et al. [13] scales trajectory-matching methods to ImageNet-1K. Chen et al. [10] proposes progressive dataset distillation, improving data quality through multiple distillation stages. In comparison, sample selection also filter out low-quality samples to obtain more informative and representative data, while sample selection obtains better interpretability.

E. Dataset Introduction

CC3M (Conceptual Captions 3M) is a large-scale multimodal dataset comprising 3.3 million images paired with textual captions, developed by Google using automated web crawling and filtering techniques. Sourced from publicly available

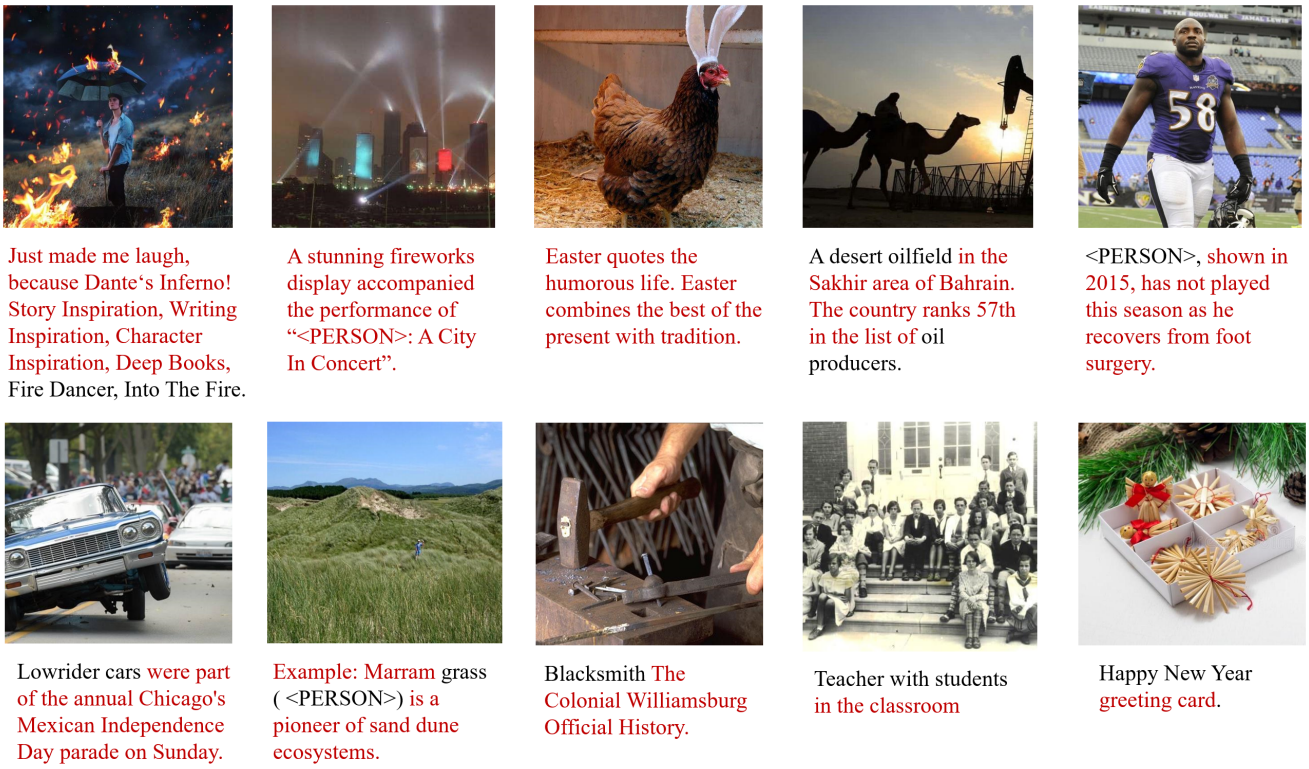

Figure 6. Visualization of noisy data pairs detected by DISSect in CC3M. The red color means the text is mismatched to the image.

web images and their contextual text descriptions, the dataset undergoes rigorous cleaning and deduplication to ensure strong alignment between visual and textual content. Recognized for its high-quality annotations and moderate scale, CC3M is commonly utilized in tasks such as image-text matching and text-to-image synthesis, particularly for training efficient multimodal models with limited computational resources.

CC12M (Conceptual Captions 12M) expands upon CC3M by scaling to 12 million image-text pairs, collected through similar automated web extraction methods but with extended coverage of diverse online sources. While it significantly enhances data diversity (encompassing broader scenes, objects, and linguistic variations), the dataset introduces more noise compared to its predecessor, necessitating advanced preprocessing or noise-tolerant training strategies. Its optimal balance between scale and variability makes CC12M a foundational resource for developing large-scale vision-language models requiring substantial training data.

YFCC15M (YFCC100M Subset-15M) is a curated subset of Flickr's YFCC100M dataset, containing 15 million images accompanied by user-generated metadata such as titles, tags, and descriptions. Derived from authentic user uploads, it reflects real-world photographic content spanning nature, culture, and daily life, characterized by high ecological validity but relatively sparse or informal textual annotations. This dataset is extensively employed in multimodal pretraining and cross-modal retrieval research, offering critical benchmarks for evaluating model performance in practical, uncurated scenarios.

F. DISSect Selected Sample Visualization

The CC3M dataset is a real-world dataset characterized by 3% to 20% noisy correspondence. In our study, the DISSect method demonstrates remarkable robustness against this real-world noise, as evidenced by its improved performance on the CC3M dataset. As shown in Fig. 6, DISSect successfully identifies several noisy samples in CC3M. We found that most of these detected noisy pairs are either completely or partially mismatched.

The causes of these mismatches are diverse. Some noisy pairs, although originating from the same web source, are not directly descriptive of each other. Other mismatches occur due to efforts to protect personal information. For the partially mismatched pairs, the text typically describes only a fragment of the image. This partial correspondence introduces ambiguity, as the same text could correspond to multiple images within the dataset.

G. Pseudo Algorithm for Temporal Ensembling Version

Algorithm 2 Pipeline of learning with DISSect.

Input: Dataset $\mathcal{D}$, Momentum β, Selection Ratio r.

Maintain $\text{CLIPScire}_{\text{hist}}$ as a dictionary.
for *each epoch* $t = T_w, \ldots, T$ **do**
  for *each batch* $\mathcal{D}_b$ *from* $\mathcal{D}$ **do**
    Forward-propagation to get features I, T.
    Look up the dictionary for $\text{CLIPScire}_{\text{hist}}$.
    Predict $\text{CLIPScore}_{\text{curr}}$ by Eq. (2).
    Compute discrepancy score δ by Eq. (3).
    Update $\text{CLIPScire}_{\text{hist}}$ by Eq. (5).
    Extract mini-batch $\mathcal{D}'_b$ from $\mathcal{D}_b$ with $\text{topk}_r \delta$.
    Calculate loss by Eq. (1) on $\mathcal{D}'_b$, back-propagation.

Output: Pre-training accelerated encoders $f_{\theta_1}, g_{\theta_2}$.

H. Ablation Studies on Different Hyper-parameters

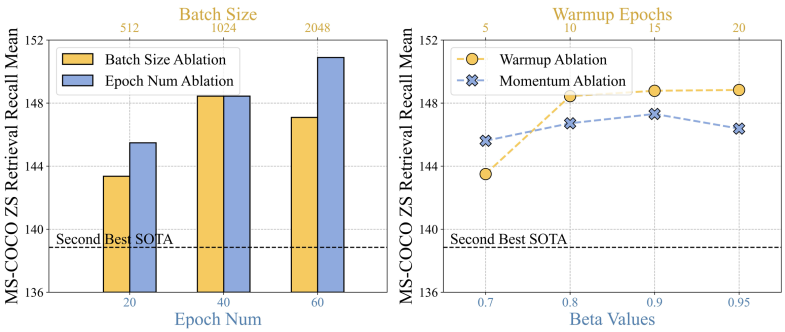

Figure 7. Ablation studies on hyper-parameters, including different batch sizes, training epochs and β values.

In Fig. 7 left, we conduct ablation studies with different batch sizes and number of epochs to demonstrate the effectiveness of DISSect under various training conditions. DISSect has proven its robustness against various extent of real-world data noise, as the well-curated CC12M and noisier YFCC15M. In Fig. 7 right, DISSect performs stably against different β values. Therefore, we recommend the momentum version with β set to a default value (0.9 in our paper) when there is no prior knowledge of given dataset.