Bi-directional Hard-Negatives Ranking Loss for Cross-Modal Video-Text Retrieval

Xiaoshuai Hao
Institute of Information Engineering
Chinese Academy of Sciences
Beijing, China
haoxiaoshuai@iie.ac.cn

Abstract

Cross-modal retrieval between videos and texts has attracted growing attentions due to the rapid emergence of videos on the web. Existing works often only use more modalities video feature embedding to measure cross-modal similarities. However, these methods ignore the design of loss function. The loss function is important for cross-modal retrieval task. To improve R@1 recall performance which is essential especially for cross-modal video-text retrieval task, we elaborately design a new bi-directional hard-negatives ranking loss (Bi-HNRL) that emphasizes on the hardest negatives. As extensive experiments on five benchmarks, the proposed loss function is much more effective on the cross-modal video-text retrieval task.

1. Introduction

Cross-modal retrieval task has attracted considerable research attention in the field of retrieval task. The rapid growth of video on the internet has made searching for video content using natural language queries a significant challenge. In such retrieval paradigm, end-user searches for unlabeled videos by ad-hoc queries described in natural language text with no visual example provided. The current dominant approach for cross-modal retrieval is to encode different modalities into a joint embedding space[2] to measure cross-modal similarities. In this paper, we focus on learning effective joint embedding models for the cross-modal video-text retrieval task.

2. Based Model

Given a video $V$ and a query text $Q$, we would like to create a pair of functions $\phi(V)$ and $\phi(Q)$ that map video data and text into a joint embedding space that respects this correspondence embryos for paired text and video that do not match should lie far apart. In this paper, we use the CE mode, which is proposed[4]. The CE model could embed the information from multiple modalities and multiple time steps of a video segment into a compact fixed-length representation.

3. Bi-direction hard-negatives ranking loss

In this section, we introduce in detail our new loss function bi-direction hard-negatives ranking loss and discuss the difference from the original rank loss in the following subsection.

In the same embedding space, embeddings for paired query sentence and video should lie close together, while embeddings for which query sentence and video do not match should lie far apart. However, many prior approaches[5, 6, 3] have utilized pairwise ranking loss for learning joint embedding between visual input and textual input. They minimize a hinge-based triplet ranking loss combining bi-directional ranking terms, in order to maximize the similarity between a video embedding and the corresponding text embedding, and at the same times, minimize the similarity to all other non-matching one. More formally, at training time, we sample a batch of video-text pairs $(Q_i, V_i)_{i \in [1, B]}$ where $B$ is the batch size. The similarity score $s_{i,i} = s(Q_i, V_i)$ between video $V_i$ and its ground truth texts $Q_i$ is greater than every possible pair of scores $s_{i,j}$ and $s_{j,i}$, where $j \neq i$ of non-matching videos and texts. The bi-directional max-margin ranking loss (Bi-MMRL) can be written as:

$$l = \frac{1}{B} \sum_{i=1}^{B} \sum_{j \neq i} \left[ \max(0, m + s_{i,j} - s_{i,i}) \right]$$

where $s_{i,j} = s(Q_i, V_j)$ is the similarity score of query sentence $Q_i$ and video $V_j$, and $m$ is the margin value for the pairwise ranking loss. This loss is composed of two sym-
metric terms, by considering $Q$ and $V$ as individual queries. In each term, the loss is defined by the sum of the violations for each negative sample.

Recently, focusing on hard-negatives has been shown to be effective in many embedding tasks [1][8][7]. Inspired by this, we propose a new bi-direction hard-negatives ranking loss function for this task that emphasizes on the hardest negatives, which significantly improves the retrieval performance. Specially, we focus on hard negatives (i.e., the negative video and query sentence closest to a positive pair) instead of summing over all negatives in our formulation. For a positive pair $s_{i,i} = s(Q_i, V_i)$, the hardest negative sample can be identified $\tilde{s}_{i,j} = \text{argmax} s(Q_i, V_j)$ and $\tilde{s}_{j,i} = \text{argmax} s(Q_j, V_i)$. The new loss function bi-directional hard-negatives ranking loss (Bi-HNRL) can be written as:

$$l = \frac{1}{B} \sum_{i=1}^{B} \sum_{j \neq i} \left[ \max (0, m + \tilde{s}_{i,j} - s_{i,i}) + \max (0, m + \tilde{s}_{j,i} - s_{i,i}) \right]$$

The main difference between the bi-directional max-margin ranking loss and the bi-directional hard-negatives ranking loss is the number of negative triplets that affect the loss at each step of stochastic gradient descent (SGD). The bi-directional max-margin ranking loss sums the violations all negatives, while the bi-directional hard-negatives ranking loss function only considers the penalty incurred by the hardest negative. This illustration shows that only focusing on the hardest negative is important to R@1 (See Fig.1). In other word, the bi-directional max-margin ranking loss is actually minimizing the mean of non-negative terms. In doing so, it is aggregating the subtle gradient signal from many samples. Therefore the gradient updates are no longer noisy and SGD may not be capable of jumping out of local minimum[1]. The new loss function bi-directional hard-negatives ranking loss function reduces the contributing terms and considers only the hardest negatives. We experimentally show that bi-directional hard-negatives rank-

4. Experiments

4.1. Datasets and Evaluation Metric

We present experiments on five benchmark datasets: MSVD, DiDeMo, ActivityNet, MSRVTT and YouCook2.

We use the standard evaluation criteria used in the End-of-End-to-End - A Video Understanding Pentathlon challenge.

4.2. Discuss of Loss function

We discuss of the effectiveness of the proposed bi-directional hard-negatives ranking loss function. The effectiveness of the proposed bi-directional hard-negatives ranking loss function is reported in Fig 2. Recall performance is compared using bi-directional max-margin ranking loss and bi-directional hard-negatives ranking loss on five benchmarks. We experimentally show that bi-directional hard-negatives ranking loss is much more effective on the cross-modal video-text retrieval task.

5. Conclusions

In this paper, we elaborately design a new bi-directional hard-negatives ranking loss (Bi-HNRL) that emphasizes on the hardest negatives, which is much more effective on the cross-modal video-text task.

References


