Proc. ViRaL Worshop @ Int. Conf. on Computer Vision (ICCV), Oct. 2021. Authors' accepted version.

Hard-Negatives or Non-Negatives? A Hard-Negative Selection Strategy for Cross-Modal Retrieval Using the Improved Marginal Ranking Loss

Damianos Galanopoulos
CERTH-ITI
Thermi-Thessaloniki, Greece
dgalanop@iti.gr

Vasileios Mezaris
CERTH-ITI
Thermi-Thessaloniki, Greece
bmezaris@iti.gr

Abstract

Cross-modal learning has gained a lot of interest recently, and many applications of it, such as image-text retrieval, cross-modal video search, or video captioning have been proposed. In this work, we deal with the cross-modal video retrieval problem. The state-of-the-art approaches are based on deep network architectures, and rely on mining hard-negative samples during training to optimize the selection of the network's parameters. Starting from a state-of-the-art cross-modal architecture that uses the improved marginal ranking loss function, we propose a simple strategy for hard-negative mining to identify which training samples are hard-negatives and which, although presently treated as hard-negatives, are likely not negative samples at all and shouldn't be treated as such. Additionally, to take full advantage of network models trained using different design choices for hard-negative mining, we examine model combination strategies, and we design a hybrid one effectively combining large numbers of trained models.

1. Introduction

Ad-hoc video retrieval, a special case of cross-modal video search, is a very challenging and important task. The goal of the task is to retrieve unlabeled video shots using only textual queries. The above scenario is directly related to a real-world video retrieval system. Users need to search for videos without prior knowledge of the available videos, without video exemplars, and without elaborate procedures for formulating their textual queries.

During the last few years, many methods have been proposed for the ad-hoc video retrieval task, mainly inspired by the directly-related TRECVID task [1]. In contrast to early solutions to this problem, which relied on pre-trained concept detectors, the majority of the recent methods aim to learn new representations for both video and text in a common feature space. They target designing powerful initial and middle-level representations for both the video and textual streams to learn the final common feature space. Moreover, recent studies show that combining the results of multiple architectures and trained models leads to improved results and more stable performance.

Inspired by the triplet loss function that is used for image retrieval [17], a common way [7] [5] [16] to train a cross-modal learning system is to emphasize on the hardest negative samples. A hard-negative is a negative sample, but at the same time, is located near to the anchor sample (i.e. the positive sample) in the feature space. Using such losses, the performance of cross-modal systems gains significant improvement [7]. Hard-negatives could be mined either offline, i.e., to compute the embeddings of all possible samples before the start of training, or online at the training stage [17], which is the most effective way. For a given anchor, the corresponding hard-negative sample is computed based on their distance in every batch.

Based on a state-of-the-art cross-modal video retrieval method, we design a method for improved hard-negative mining. We focus on extracting actual hard-negative samples by identifying the candidate hard-negatives that are nevertheless semantically closeby to the anchor, thus should not be treated as negatives at all. Adding this new procedure to the overall method, and varying a simple parameter that controls how "semantically closeby" to the anchor a hard-negative is allowed to be, we end up with a multitude of trained models. These models should be combined in an efficient way to boost further the performance of our system. For this, we examine different strategies on how we can effectively combine multiple trained models to improve performance.

The contribution of this paper is twofold. Firstly, a method for hard-negative mining is introduced, evaluated, and compared with the baseline improved marginal ranking loss [7]. Secondly, we introduce a strategy for effectively combining multiple trained modes.

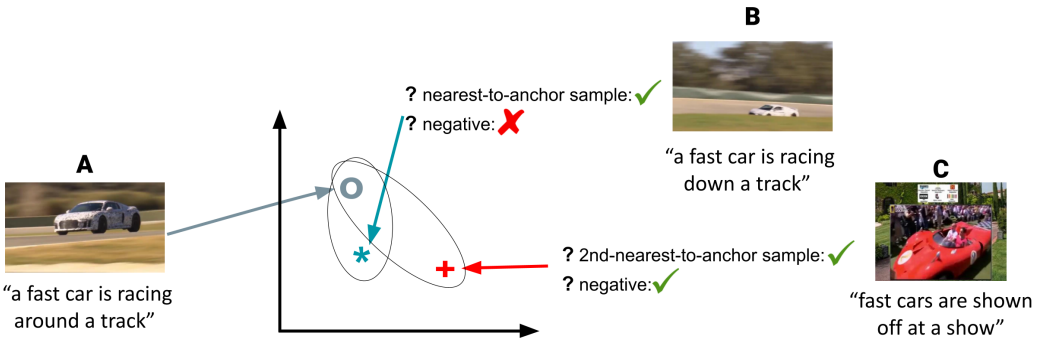

Figure 1. Given as anchor the video-caption sample A, which one of B, C should be used as a hard-negative sample during training? B, which is the nearest sample to A, but in fact would be a good match to a query based on the caption of A? Or C, which is a bit more distant from A in the feature space that is being learned, but is a true negative sample?

2. Related work

Early approaches [13] [14] to cross-modal retrieval relied on representing different modalities, e.g., video, text, or audio, into a predefined set of concepts. Contrarily, most of the recently-proposed works focus on learning jointly visual- and textual-embedding spaces. In [10] a combination of three different textual encodings leads to better overall textual encoding, and in combination with powerful video features lead to efficient video-text matching. [5] and [8] utilize two similar branches of multilevel encodings, both for the video and textual streams. In [18] multiple and diverse representations by combining global and local features using multi-head attention are used to deal with the polysemous instances in the video-text retrieval problem.

More recent approaches also try to combine multiple architectures for video-text retrieval for improved embedding space learning. A hierarchical graph-based method is utilized in [3] to decompose video-text matching into global-to-local levels. In contrast, [11] proposed a model, which combines multiple simple encoders, pre-trained and fixed, in order to create multiple embedding spaces and to combine them using multi-loss learning. Similarly, in [6] a hybrid space learning that effectively combines a latent space with a concept space is proposed.

For the majority [3] [6] [5] of video-text matching and retrieval methods when it comes to the learning procedure, the triplet ranking loss [9] and its improved versions [7] are extensively used. A modified version of the pairwise ranking loss is proposed in [16], and a weighed ranking loss emphasizing on the hard-negatives is designed. The majority of the methods focus on semi-hard negatives, e.g., the negatives inside a mini-batch, instead of mining the hard-negatives in the entire training dataset. Inspired by these works, we focus on finding the real hard-negatives inside a mini-batch to discard potentially positive samples from being treated as hard-negatives.

3. Baseline method

As a starting point of this work, we utilize the attention-based dual encoding network presented in [8]. This network is trained to transform an input video-caption sample (v, c) into a new joint embedding space $\phi(\cdot)$. The network consists of two similar sub-networks, one for the video stream and one for the textual one. Each sub-network consists of multiple levels of encoding, i.e. using mean-pooling, bi-GRU, and CNN, layers. Following the state-of-the-art approach [5] [7] [8], the improved marginal ranking loss [7] is used to train the entire network.

Following [8], where the combination of multiple models is shown to lead to improved performance, we train 24 different models by modifying the following parameters: two positions in the architecture are considered for inserting the attention mechanism (textual or visual stream) $\times$ two textual encodings are used (BERT, W2V+BERT) $\times$ two optimizers (Adam, RMSprop) $\times$ three learning rates $(1 * 10^{-4}, 5 * 10^{-5}, 1 * 10^{-5})$. The resulting 24 models are combined in a late fusion scheme (i.e, averaging a given sample's ranking in the 24 resulting ranking lists).

4. Hard-negative mining

The improved marginal ranking loss introduced in [7] and extensively used in video retrieval approaches, among others in [8] and [5], emphasizes on the hard-negative samples in order to learn to maximize the similarity between textual and video embeddings. Given a video-caption sample (v, c), the improved marginal loss is defined as follows:

$$\mathcal{L}(v,c) = max(0, S(v',c) - S(v,c)) + max(0, S(v,c') - S(v,c)) \quad (1)$$

where $S(v, c)$ is the similarity between two items, v' and c' are the hardest negatives of c and v respectively. However, when it comes to unlabeled data, such as video-caption

samples, it is difficult to find hard-negative samples since no labels or classes exist in order to identify samples from the same or other classes. At the training stage, and in every batch during the network's training, for a given anchor sample (v, c), all other samples in the same batch are considered negatives. The most common approach to mine the hardest negative is by selecting the sample nearest to the anchor (i.e., the most similar sample in the embedding space [17]). But is this sample actually a hard-negative? This is the question we try to answer. We will examine if a selected sample is actually a hard-negative or a positive one.

An illustration of the above problem is shown in 1. Considering as anchor the video-caption sample A with the caption *"a fast car is racing around a track"*, the closest sample B, i.e., with the highest similarity score, has the caption *"a fast car is racing down a track"* and the second closest sample C is captioned as *"fast cars are shown off at a show"*. Following the common hard-negative mining procedure, B would be selected as hard-negative, but obviously it is a positive sample. On the other hand, C, the second nearest to A sample, is similar to A but, clearly is a negative sample and should be selected as hard-negative instead of B.

For this reason, we designed a strategy to exclude potentially-positive samples. First, we randomly split the training dataset into batches, similarly to the standard training procedure. In each batch we compute the cosine similarity score $S^{bert}_{i,j}$, based on the initial BERT [4] representations of captions, between all possible captions (c_i, c_j) inside the batch. By collecting all these scores, we can compute a threshold value p for which $x\%$ of the $S^{bert}_{i,j}$ similarities are higher than p. At the training stage, for an anchor (v_i, c_i), every sample (v_j, c_j) (within the batch) with $S^{bert}_{i,j} > p$ is excluded from the negatives, while every other sample is labeled as negative. Finally, as hard-negative, the negative sample with the highest $S^{bert}_{i,j}$ is selected.

5. Fusion strategies

As discussed in Section 3, the combination of multiple models is known to boost the performance. Using the dual encoding network of [8], we end up with 24 trained models. Moreover, using the proposed strategy for hard-negative mining and by varying parameter x, the number of models can be quickly increased. An efficient way to combine all these models is needed to achieve the optimal result. In this section, we discuss the strategies we designed to combine all these models in order to finally retrieve the most related videos for a given textual query.

Every trained model q, for a given query que, results in a ranking list of the n most relevant videos, $R^q_{que} = \{v_1, v_2, \ldots, v_n\}$. We study the combination of multiple ranking lists in order to find the best-performing strategy when a plethora of ranking lists is used. Three different strategies are examined.

- *AVG*: For every video, its rankings in all $R^q_{que}, q = 1, \ldots, Q$ are averaged to calculate its final ranking: $rank^{Avg}_v = \frac{1}{Q}\sum_{q=1}^{Q} rank^q_v$, where Q is the number of the ranking lists, $rank^q_v$ is the video ranking in the ranking list R^q_{que}. This is the approach of [8].

- *MAX*: The final video ranking is calculated as the maximum ranking across all ranking lists Q. $rank^{Max}_v = \max_q \{rank^q_v\}$

The assumption behind the *AVG* approach is that every model we train is a well-performing one, thus treating them equally and averaging the rankings for a given video is a meaningful way of combining them. Contrarily, the assumption behind the *MAX* approach is that our models may not be very accurate, but they are most likely correct in identifying true positives at least at the very top of the ranking lists they produce. Thus, if a video appears very high in the ranking list generated by at least one model, we trust this video to be a good answer to the query.

As neither of these two assumptions seems perfectly plausible, we propose a hybrid strategy where, for a retrieved video, we identify the Q' ranking lists where the video is ranked the highest among the Q in total ranking lists, and we average its top-Q' rankings. This average is calculated for every video and is used for ultimately ranking the retrieved videos in descending order. I.e.,

- *Hybrid*: The top-Q' video rankings across all ranking lists are used to calculate the final video ranking. $rank^{Hyb}_v = \frac{1}{Q'}\sum_{q=1}^{Q'} rank^q_v$, where $1 \leq Q' \leq Q$.

  If $Q' = 1$ then $rank^{Hyb}_v = rank^{Max}_v$ and if $Q' = Q$ then $rank^{Hyb}_v = rank^{Avg}_v$.

6. Experimental results

6.1. Experimental setup

To train our networks we use a combination of four large-scale video caption datasets: MSR-VTTT [20], TGIF [12], ActivityNet [2] and Vatex [19]. We evaluate the networks' performance on the official TRECVID AVS dataset for 2019 and 2020, i.e., the V3C1 dataset. The evaluation measure we use is the mean extended inferred average precision (MxinfAP). As initial frame representations, a ResNet-152 (trained on the ImageNet-11k dataset) is used. Regarding the textual parts, two different word embeddings are utilized: i) the Word2Vec model [15] and ii) BERT [4].

6.2. Results

Table 1 presents the results when the $x = 1\%$ and $x = 2\%$ of the samples are excluded from the hard-negative mining because of being treated as potentially positive samples, compared to the baseline, using the examined model

Table 1. Results, in MxinfAP, of the combination of multiple models and different setups. The baseline hard-negative mining strategy is compared with the proposed hard-negative mining one, with $x = 1\%$ and $x = 2\%$ exclusion, for three fusion strategies. The last row shows the results of combining all models trained using different hard-negative mining methods. The best results among the three hard-negative mining approaches for a given fusion strategy and test dataset are underlined, while the results in bold are the overall best.

	AVG		MAX		Hybrid, Q'=10	
	AVS19	AVS20	AVS19	AVS20	AVS19	AVS20
Baseline hard-negative mining (24 models)	0.1483	0.2300	0.1414	0.2182	0.1480	0.2300
Proposed hard-negative mining, $x = 1\%$ (24 models)	<u>0.1492</u>	<u>0.2303</u>	<u>0.1428</u>	<u>0.2212</u>	<u>0.1492</u>	<u>0.2304</u>
Proposed hard-negative mining, $x = 2\%$ (24 models)	0.1480	0.2264	0.1419	0.2139	0.1482	0.2266
Combination of all models (72 models)	0.1493	0.2293	0.1425	0.2179	**0.1537**	**0.2416**

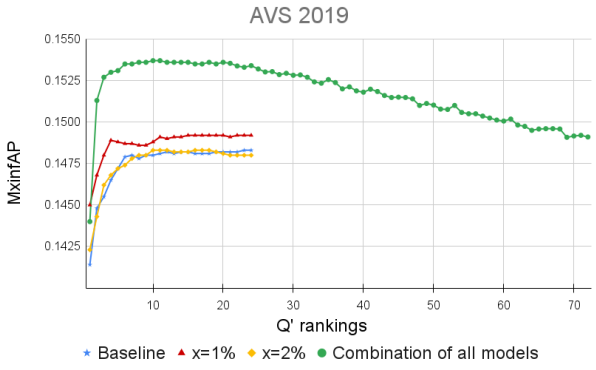

Figure 2. Results on the AVS19 dataset, in MxinfAP, for the Hybrid fusion strategy and different values of Q'.

Figure 3. Results on the AVS20 dataset, in MxinfAP, for the Hybrid fusion strategy and different values of Q'.

combination strategies. As baseline we used the improved marginal ranking loss [7] with the standard hard-negative mining procedure (i.e., for an anchor sample all other samples in a batch are considered as negatives). The performance of the combination of all models, i.e., the baseline and the proposed hard-negative mining with $x = 1\%$ and $x = 2\%$, shown in the last row of Table 1. This combination consists of 72 different models.

Using the *AVG* fusion strategy and excluding the $x = 1\%$ of the video-caption samples from the hard-negative mining, our network improves its performance, i.e., from MxinfAP 0.1483 to 0.1492 and 0.2300 to 0.2303 on AVS19 and AVS20, respectively. When we increase the exclusion parameter x to 2%, the performance slightly decreases to 0.1480 from 0.1483 and 0.2264 from 0.2300 on AVS19 and AVS20, respectively. When we combine all models using the *AVG* approach, the overall performance slightly improves on AVS19, but slightly decreases on AVS20. The *MAX* combination approach performs analogous to the *AVG*, i.e., $x = 1\%$ performs better than the baseline, but achieves lower results than *AVG*. Generally, setting $x = 1\%$ we can see a small but consistent improvement compared with the baseline, following any fusion strategy.

Regarding the Hybrid fusion strategy, we can see in Table 1, the Hybrid fusion performs considerably better when a plethora of models (72) are available. Setting $Q' = 10$, it achieves MxinfAP 0.1537 compared to 0.1492 on the AVS19 dataset and 0.2416 from 0.2304 on the AVS20 dataset. Also, by examining Figures 2 and 3, we can see that when we combine multiple models, the performance increases immediately (when $Q' \geq 2$), and there is a sweet spot $5 < Q' \leq 25$, where the performance is generally stable and is maximized. As more and more models are added, the performance then starts to decrease until it becomes equal to the score obtained by the *AVG* strategy.

7. Conclusions

We examined a new strategy for hard-negative mining to improve the performance of a cross-modal video retrieval network. We focus on excluding positive samples from being wrongfully utilized as hard-negatives. Moreover, this strategy enables generating a larger number of trained models; for this, we also proposed a hybrid strategy for model combination. From the experimental results, we conclude that the new hard-negative mining strategy is meaningful, and together with the hybrid model combination strategy boosts the video retrieval performance.

Acknowledgment. This work was supported by the EU Horizon 2020 programme under grant agreement 832921 (MIRROR).

References

[1] George Awad, Asad A Butt, Keith Curtis, Jonathan Fiscus, et al. Trecvid 2020: comprehensive campaign for evaluating video retrieval tasks across multiple application domains. In *Proceedings of TRECVID 2020*. NIST, USA, 2020.

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015.

[3] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10638–10647, 2020.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *arXiv preprint arXiv:1810.04805*, 2018.

[5] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9346–9355, 2019.

[6] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, February 2021.

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[8] Damianos Galanopoulos and Vasileios Mezaris. Attention mechanisms, signal encodings and fusion strategies for improved Ad-hoc video search with dual encoding networks. In *Proceedings of the ACM International Conference on Multimedia Retrieval*, (ICMR '20). ACM, 2020.

[9] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

[10] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. W2VV++ fully deep learning for ad-hoc video search. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1786–1794, 2019.

[11] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. Sea: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 2020.

[12] Yuncheng Li, Yale Song, Liangliang Cao, et al. TGIF: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650, 2016.

[13] Yi-Jie Lu, Hao Zhang, Maaike de Boer, and Chong-Wah Ngo. Event detection with zero example: Select the right and suppress the wrong concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, ICMR '16, page 127–134, New York, NY, USA, 2016. ACM.

[14] Foteini Markatopoulou, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. Query and keyframe representations for ad-hoc video search. In *Proceedings of the 2017 ACM International Conference on Multimedia Retrieval*, ICMR '17, pages 407–411. ACM, 2017.

[15] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, Workshop Track Proceedings*, ICLR '13, 2013.

[16] 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 *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, page 19–27, New York, NY, USA, 2018. ACM.

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[18] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1979–1988, 2019.

[19] Xin Wang, Jiawei Wu, Junkun Chen, et al. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4581–4591, 2019.

[20] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016.