Relevance-based Margin for Contrastively-trained Video Retrieval Models

Alex Falcon
falcon.alex@spes.uniud.it
Fondazione Bruno Kessler
Trento, Italy
University of Udine
Udine, Italy

Swathikiran Sudhakaran
swathikirans@gmail.com
Samsung AI Center Cambridge
Cambridge, United Kingdom

Giuseppe Serra
giuseppe.serra@uniud.it
University of Udine
Udine, Italy

Sergio Escalera
sergio@maia.ub.es
University of Barcelona and
Computer Vision Center
Barcelona, Spain

Oswald Lanz
lanz@inf.unibz.it
Free University of Bozen-Bolzano
Bolzano, Italy

ABSTRACT

Video retrieval using natural language queries has attracted increasing interest due to its relevance in real-world applications, from intelligent access in private media galleries to web-scale video search. Learning the cross-similarity of video and text in a joint embedding space is the dominant approach. To do so, a contrastive loss is usually employed because it organizes the embedding space by putting similar items close and dissimilar items far. This framework leads to competitive recall rates, as they solely focus on the rank of the groundtruth items. Yet, assessing the quality of the ranking list is of utmost importance when considering intelligent retrieval systems, since multiple items may share similar semantics, hence a high relevance. Moreover, the aforementioned framework uses a fixed margin to separate similar and dissimilar items, treating all non-groundtruth items as equally irrelevant. In this paper we propose to use a variable margin: we argue that varying the margin used during training based on how much relevant an item is to a given query, i.e. a relevance-based margin, easily improves the quality of the ranking lists measured through nDCG and mAP. We demonstrate the advantages of our technique using different models on EPIC-Kitchens-100 and YouCook2. We show that even if we carefully tuned the fixed margin, our technique (which does not have the margin as a hyper-parameter) would still achieve better performance. Finally, extensive ablation studies and qualitative analysis support the robustness of our approach. Code will be released at https://github.com/aranciokov/RelevanceMargin-ICMR22.

KEYWORDS
deep learning, cross-modal retrieval, video retrieval, relevance

ACM Reference Format:
Alex Falcon, Swathikiran Sudhakaran, Giuseppe Serra, Sergio Escalera, and Oswald Lanz. 2022. Relevance-based Margin for Contrastively-trained Video Retrieval Models. In Proceedings of the 2022 International Conference on Multimedia Retrieval (ICMR '22), June 27–30, 2022, Newark, NJ, USA. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3512527.3531395

1 INTRODUCTION

With the rapid growth of digital media shared on the web it becomes increasingly important for real-world applications to offer flexible, user friendly modalities to access media content at scale. Google video search for example, translates a natural language query into a ranked list of content-related videos from the web. Natural free form, unrestricted language enables a user to express the fine-grained details in an articulated query, and each user can do so with its own expressivity. Thus, a same retrieval response can be triggered with syntactically different but semantically coherent queries. This poses significant challenges to the current state of the art in cross-modal retrieval research.

Recent approaches which deal with cross-modal video retrieval aim at learning a joint embedding space [8, 13, 18, 52] by means of contrastive losses [24, 43, 46, 48], which put the associations available in the dataset (e.g. a video and its natural language description) as close as possible while enforcing a separation margin to all the other items (see lower left of Fig. 1). During inference, the ranking list for a given query is produced by computing similarity scores with respect to all the items by means of, e.g. the dot product or the cosine similarity. By measuring the performance of the video retrieval system with rank-unaware metrics, such as recall rates, increasingly better solutions to this problem were proposed. In fact, contrastive losses synergize well with recall rates, given how they maximize the similarity of the associated items. But during training they do not make any distinction between items which are highly relevant and items which are only partially or completely irrelevant to a given query. For example, if a query is about ‘how to cook a pizza’, then videos which depict how to ‘bake a pizza’, ‘cook pasta’, or ‘knead dough’ are all treated the same way, although they can
We propose to move away from such a paradigm and adopt a video retrieval consists in visual-language datasets only providing computed by multiple models. Differently from them, we propose to relevance the item is to the query, as illustrated in Fig. 1. By doing

```
margin proportional to how relevance-based margin. To do so, we leverage the relevance function\[ R \]
```

Figure 1: Training a model for text-video retrieval by employing a contrastive loss which uses a fixed margin \( \Delta \) (lower left) treats semantically equivalent descriptions which do not appear as groundtruth pairs in the dataset as equally irrelevant. We propose to move away from such a paradigm and adopt a relevance-based margin (lower right), i.e. a margin which is proportional to the relevance \( R \) (see Sec. 3.1).

be more or less semantically close to the query. Furthermore, one of the reasons which limited the usage of rank-aware metrics in video retrieval consists in visual-language datasets only providing the visual contents and textual annotations (obtained manually [55, 61] or automatically [44]). Due to the absence of relevance grades, rank-aware metrics (e.g. nDCG) are difficult to adopt. Recently, this problem was partially alleviated by the introduction of a relevance function [15] which, to avoid a costly manual annotation step, is defined in terms of the captions already available in the dataset.

To give the model awareness of the semantical differences between items and queries during training, we free the margin from its stillness. Several solutions for non-fixed margins were proposed in previous literature, such as using multiple margins (e.g. [11]) or adaptive solutions. In particular, [49] implemented a schedule for the margin value which gradually incorporates inter-category correlations and information about the structure of the embedding space. Recently, for video retrieval [25] proposed an adaptive margin proportional to the similarity of item and query as computed by multiple models. Differently from them, we propose to inject semantic knowledge into the training process by means of a relevance-based margin. To do so, we leverage the relevance function detailed in [15], so that the margin is proportional to how relevant the item is to the query, as illustrated in Fig. 1. By doing so, we effectively discard one hyper-parameter to tune. Moreover, even by performing an expensive search for it, the results are still suboptimal when compared to the proposed relevance-based margin. We give empirical evidence that the proposed technique makes it possible to easily improve the quality of the ranking lists, measured through Normalized Discounted Cumulative Gain (nDCG) and Mean Average Precision (mAP). We use three different and increasingly more complex models (MME from [54], JPoSE [54], and HGR [8]) on two datasets (EPIC-Kitchens-100 [15] and YouCook2 [61]). Furthermore, we perform several ablations to study how it interacts with multiple video modalities (motion, appearance, audio) and with both cross-modal and within-modal losses.

We organize the paper as follows. In Section 2 we review related works, including vision and language tasks, main techniques and losses used to deal with text-video retrieval, and optimization of retrieval metrics such as the nDCG. Then, we formally describe the proposed technique in Sec. 3, in terms of the relevance function and how we apply it to a typical contrastive loss setting. In Sec. 4 we perform multiple experiments to prove the strength of the relevance-based margin. Finally, in Sec. 5 we conclude the paper.

2 RELATED WORKS

Vision and Language. In recent years, deep learning brought several advancements in multiple tasks dealing with vision and language, such as question answering [1, 2, 28, 32], retrieval [8, 18, 33, 58], and captioning [19, 35, 37, 50]. Given that vast amounts of data can be scraped from the web, many works perform a joint vision and language pretraining [10, 40, 51, 60] by optimizing vision-text proxy tasks. Recently, a line of research uses natural language supervision such as captioning [17] or alignment [30] objectives to pretrain visual models. While in both cases they achieve competitive and state-of-the-art results on downstream tasks, these methods are data hungry and expensive to train, making them impractical from a computational point of view.

Text-Video Retrieval. Multiple techniques were proposed to learn a representation for the input data while capturing multimodal interactions. [21, 42, 52] explore multimodal fusion techniques to fuse all the information extracted from a video using multiple pretrained ‘experts’. While these methods focus on the addition of video-side information, a supervisory signal can also be obtained by looking with more detail at the text. [8] create a semantic role graph of the caption and aligns to each node a learned representation of the clip-level descriptor. [54] extract verbs and nouns from the caption and uses them to learn Part-of-Speech-specific embedding spaces. [47] introduce a generative cross-captioning task, using the batched videos as a support set. Recently [13] distil information from multiple pretrained text experts. A different trend involves heavy pretraining steps [4, 20, 34, 41], followed by fine-tuning for downstream tasks. Moreover, the addition of image-text datasets as part of the pretraining step, showed significant improvements when dealing with video-related tasks [4, 34]. While these methods achieve impressive results, they rely heavily on the data, are expensive to train, and are not designed for the nature of the problem.

Due to the unavailability of groundtruth relevance values which can inform about the optimal ranking list to a given query, the video
retrieval community focused on rank-unaware metrics such as the recall rates or the median rank. Contrastive losses greatly improve these metrics since they reduce the distance between the visual descriptor and the linguistic one and thus increase its similarity, making it possible to retrieve it before the negative descriptors. But multiple descriptions can be equally or partially relevant for the same video (and vice versa), thus more complex and rich metrics, such as the nDCG, are needed to accurately evaluate a retrieval system [53]. To do so, a way to determine how relevant an item is to a query must be available. To avoid the need for manual and costly annotation, [15] proposes to use a relevance function defined in terms of the noun and verb classes present in the caption (more details in Sec. 3.1).

Learning a joint embedding space. Common approaches for text-video retrieval learn a joint embedding space by means of a contrastive loss [24, 48] which, during training, puts semantically similar items (e.g. a video and a caption describing its contents) closer in the embedding space, while dissimilar items are pushed away. While groundtruth associations (i.e. positive pairs, such as a video and its caption) are known from the dataset, the negative examples (such as a different video) have to be sampled, or ‘mined’, given that the amount of possible tuples scales exponentially with the dataset size, e.g. cubically with triplets. Multiple techniques have been proposed including: offline mining, which randomly samples a fixed number of tuples and repeats the process multiple times during training; online mining, which uses the negatives inside the mini-batch by considering all the non-groundtruth pairs, or only hard [26, 56] or semi-hard negatives [48]. Recent research also found relevant signal while mining positive samples, e.g. easy [57] or hard positives [26]. In our paper, we focus on triplets as they are a popular margin-based contrastive loss, but it can be extended to other techniques, e.g. to quadruplets [9]. Moreover, we experiment with two different and opposite techniques: offline mining with random sampling and online mining with hard negatives, and show the advantages of the relevance-based margin in both cases.

Margin in contrastive losses. Most of the approaches involving contrastive losses are based on maximum-margin losses (e.g. [24]). Although the margin is usually fixed, variable or adaptive solutions for it have been explored in different fields. For person re-identification, [11] suggest using two different (but fixed) margins for inter- and intra-class constraints, whereas [59] propose to monotonically increase the margin during the training process. [27] use a ‘soft margin’ to improve recommender systems, that is they remove the fixed margin and use a (soft version of) the distance between positive and negative pairs as the loss. [39] augment the bidirectional contrastive loss by also summing the margin to the loss objective, to optimize it during the training process. For text-image retrieval, [49] propose a scheduled adaptive margin which starts from a fixed value and gradually changes during the training process both to integrate inter-category similarity-based correlations and to preserve the category clusters formed during the initial phases of the training. Recently, for cross-modal video retrieval [25] proposed an adaptive margin proportional to the similarity of the representations computed for the negative pair, both in terms of ‘static’ (pretrained, frozen) models, which provide initial supervision, and ‘dynamic’ (trained with the task) models, which provide supervision in later stages of the training. Differently from all these works, we propose a margin which is proportional to the relevance value of the queries involved in the triplet, effectively using the semantic knowledge during training.

Optimization of nDCG. Considering that visual-textual datasets usually lack relevance grades, rank-unaware metrics are one of the preferred ways to measure progress in the video retrieval community. Yet given a video, multiple captions can be used to describe its contents. To capture the difference in the ranking list when binary relevance (i.e. a caption is either relevant or irrelevant to a video) is considered, mAP is preferred to the recall rates. Furthermore, finer-grained relevance grades could be also available (i.e. a caption can be relevant to a video to some degree), in which case the DCG (or its normalized version, the nDCG) is chosen. But, optimizing these metrics during training clashes with gradient-based optimization methods because ranks are not differentiable with respect to the learnable parameters, e.g. the nDCG of a list of items to a given query is normalized using the optimal ranking list, which is computed by sorting with respect to the relevance values.

Surrogate losses are used to partially address this problem, which can be categorized into: pointwise (e.g. regression loss [12]), which compare predicted and optimal rank of one item at a time; pairwise (e.g. RankNet [7]), which deal with pairs of items and relative ordering; listwise approaches (e.g. LambdaRank [6]), which work on full list of items. Note that the triplet loss [48] can be seen as a ‘triplet-wise’ surrogate loss. Since these surrogate losses are loosely connected to downstream metrics, there is also an active research field which directly optimizes retrieval metrics by deriving a relaxation of the sorting operator which has well-defined gradients, e.g. [5, 14, 22].

Considering its widespread usage for video retrieval, we consider the triplet loss an optimal candidate for our relevance-based margin, and show it can lead to higher quality ranking lists.

3 RELEVANCE-BASED MARGIN

In Sec. 3.1 we define the relevance function $R$ and the metrics used during evaluation. In Sec. 3.2 we describe how we change the margin in the contrastive loss to make it dependent on $R$. Finally, Sec. 3.3 details the three methods on which we test our technique.

3.1 Semantic classes and relevance

Given a video clip, multiple natural language descriptions may fully capture its visual contents, and vice versa. Hence, if a user looks for videos about ‘cooking a pizza’, an intelligent video retrieval system should retrieve all the videos which show how to cook a pizza, and show them all before (i.e. rank them higher than) those that show the baking of a ‘focaccia’. Similarly, videos about ‘fried potatoes’ should be ranked even lower, given how dissimilar they are when compared to the user query. As a consequence, the automatic evaluation of the quality of a ranking list requires a function which considers ‘focaccia’ more relevant than ‘potatoes’ when compared with ‘pizza’, as well as the cooking technique (‘bake’ versus ‘fry’). To avoid the need for costly manual annotation which requires human assessments using a predefined set of grades, [15]
introduces a relevance function \( R \) defined as:
\[
R(x_i, x_j) = \frac{1}{2} \left( \frac{|x_i^Y \cap x_j^Y|}{|x_i^Y \cup x_j^Y|} + \frac{|x_i^N \cap x_j^N|}{|x_i^N \cup x_j^N|} \right)
\]
where \( x_i^Y \) and \( x_i^N \) denote the sets of verb and noun classes found in the \( k \)-th caption. This can be extended to videos by considering the associated description. By defining the relevance as in Eq. 1, \( x_i \) is highly relevant to \( x_j \) if they share similar noun and verb classes. We refer to ‘classes’ because we do not want to consider synonyms (e.g. ‘pick up’ and ‘take’, or ‘drop’ and ‘put down’) as different items which need to be separated, hence each class will contain tokens with a similar meaning. In some datasets, this class knowledge may be already available, but several other datasets do not provide it. To automatically compute them, a pipeline made of a PoS-tagger (e.g. with spaCy), followed by WordNet [45] and the Leuk algorithm [36] can be used, as in [53].

To evaluate a video retrieval system, we use two metrics which are commonly used in Information Retrieval, which are the Mean Average Precision (mAP) [3]) and the Normalized Discounted Cumulative Gain (nDCG) [29], as recently proposed in [53]. The mAP is defined as the mean of the Average Precision (AP) with respect to all the queries. For a given query \( q \), AP can be defined as:
\[
AP(q) = \frac{\sum_{k=1}^{N} P(k) \cdot r(k)}{N_R}
\]
where \( N \) is the number of items (both relevant and irrelevant) in the ranking list, \( P(k) \) is the Precision at \( k \) [3], \( r(k) \) is an indicator function which tells whether the \( k \)-th item is relevant or not, and \( N_R \) is the total number of relevant items. The mAP allows to grasp with a single number the area under the Precision-Recall curve. But this metric requires binary relevance values, thereby requiring the introduction of a threshold below which items are considered irrelevant (and relevant otherwise). For mAP, we consider \( k \) to be relevant to \( q \) only when \( R(q, x_k) = 1 \) as is done in [15] (hence, for mAP \( N_R = |\{x_i | R(q, x_i) = 1\})|). On the other hand, nDCG makes use of non-binary relevance values, allowing it to grasp finer details (and errors) of the ranking list. Given a query \( q \) and a list of items \( K = \{x_i\} \), it is defined as
\[
nDCG(q, K) = \frac{DCG(q, K)}{IDCG(q, K)}
\]
where IDCG is the optimal DCG value obtained when the ranking list follows a descending order of relevance values. We define DCG as in [15, 29]:
\[
DCG(q, K) = \frac{\sum_{k=1}^{N_R} \frac{R(q, x_k)}{\log_2(k+1)}}{\sum_{k=1}^{N_R} \frac{\text{ideal}R(q, x_k)}{\log_2(k+1)}}
\]
where \( x_k \) is the \( k \)-th item in the list \( K \), and we only consider the first \( N_R \) items in the ranking list. Note that \( N_R = |\{x_i | R(q, x_i) > 0\}|.

3.2 Contrastive loss with relevance-based margin

To learn a joint text-video embedding space, various contrastive (or ranking) losses have been proposed (see Sec. 2). In our work we consider a contrastive term based on the triplet loss defined as:
\[
\mathcal{L} = [m + s(a, n) - s(a, p)]_+
\]
where \([·]_+ = \max(0, ·)\), \( m \) is interpreted as a separation margin, \( s(·, ·) \) is a similarity metric (e.g. cosine similarity), whereas \( a, n \), and \( p \) represent respectively the embedding of the anchor, negative, and positive item. Eq. 5 provides a positive loss when the margin \( m \) between the positive pair \((a, p)\) and the negative one \((a, n)\) is violated, i.e. \( s(a, p) - s(a, n) < m \). The loss may be cross-modal, i.e. \( n, p \) from one modality (e.g. video) and \( a \) from the opposite one (e.g. text), or within-modal, i.e. \( a, p, n \) are all from the same modality. Furthermore, the optimal \( m \) is not known beforehand and should be treated as an hyper-parameter which can affect the performance. Thus, it should be tuned on the validation set.

During training, all the items which are not from the positive pair \((a, p)\) are pushed away until they are separated by a margin of \( m \), as shown in Fig. 1. Although effective and widely used in the literature, Eq. 5 ignores that multiple items may be completely or partially relevant to the same query, and treats all the items which are not from the groundtruth pair as equally irrelevant. Thus the retrieval system might not be able to distinguish between the many relevance levels which can exist between an item and a query.

To address this, we propose a relevance-based margin instead of a fixed margin. In our work, we aim at defining \( m \) in terms of the relevance function \( R \). In particular, we update Eq. 5 as follows:
\[
\mathcal{L} = [\Delta_{a, p, n} + s(a, n) - s(a, p)]_+
\]
where:
\[
\Delta_{a, p, n} = R(a, p) - R(a, n) = 1 - R(a, n)
\]
since we consider the groundtruth pair to be maximally relevant, i.e. \( R(a, p) = 1 \). The relevance-based margin keeps \( \mathcal{L} \) positive until \( s(a, p) \) and \( s(a, n) \) are separated by a margin which is proportional to their relevance values, thus separating irrelevant items more than those which have a positive relevance. This is illustrated in Fig. 1 on the lower right. Note that this term is not bound to the network state and can thus be applied both to offline and online mining techniques.

3.3 Methods

Given a dataset \( D = \{(a_i, q_i)\} \) of video-caption pairs, we strive to learn optimal weights for two embedding functions \( f : \mathbb{R}^{f_v} \rightarrow \mathbb{R}^d \) and \( g : \mathbb{R}^{f_t} \rightarrow \mathbb{R}^d \) such that \( f(a_i) \) and \( g(q_i) \) are close in the \( d \)-dimensional joint embedding space. Here \( f_v \) and \( f_t \) represent the dimension of the video and caption descriptors. To parameterize \( f \) and \( g \) we consider the following methods: MME is a baseline from [54] which learns one embedding function for each of the two modalities, video and text. In JPoSE [54], the captions are processed in order to obtain a single sentence-level descriptor and multiple descriptors restricted to specific Part-of-Speech (PoS) tags, e.g. nouns and verbs. Then, two functions are learned for each of these embedding spaces using both cross-modal and intra-modal contrastive terms for the sentence-level, as well as for the PoS-level. HGR [8] structures the learning at multiple levels (global event, local actions, and local entities) which are obtained by computing a semantic role graph for each of the captions. Then a graph convolutional network is used to learn compositional semantics of the caption based on the local components, i.e. full sentence, verbs, and noun phrases.
We choose these three methods because they provide incrementally structured approaches to deal with video and language data, starting from a simpler MLP-based network to a graph-based approach. Moreover, JPoSE represents the state-of-the-art for EPIC-Kitchens-100 (measured through nDCG and mAP), which is the main dataset under consideration. Finally, by selecting them we can validate our approach on both offline (MME and JPoSE) and online (HGR) mining techniques. We thus proceed to show the generality and effectiveness of the proposed relevance-based margin by empirically validating on two different datasets.

4 EXPERIMENTS

After the introduction of the experimental setting in Sec. 4.1, we show in Sec. 4.2 how the proposed relevance-based margin helps to achieve better nDCG and mAP on EPIC-Kitchens-100 and YouCook2. Then, in Sec. 4.3 we perform several ablation studies. First we show that even by carefully tuning the fixed margin, the proposed technique consistently achieves better results without the need to tune it. Secondly, we also evaluate its robustness by ablating the loss function and the modalities used in JPoSE. Finally in Sec. 4.4 we analyze the distribution of the margin values during training and some video-to-text examples from the testing set.

4.1 Experimental setting

Datasets. We focus our experimental setting on two challenging video and language datasets: the recently released EPIC-Kitchens-100 [15] and YouCook2 [61]. For the retrieval challenge, the former comprises 67217 egocentric clips for training and 9668 for evaluation. It is also the largest dataset for video retrieval in the egocentric setting. Moreover, it also provides semantic annotations for each of the captions, by covering 300 noun and 97 verb classes. The latter provides a lower amount of training clips (10337) but still offers a challenging evaluation set with 3492 clips. While semantic annotations are not available for YouCook2 they can be computed using WordNet and the Lesk algorithm, as described in Sec. 3.1. Furthermore, as both EPIC-Kitchens-100 and YouCook2 share the kitchens domain, the class knowledge of the former can also be used for the latter [53].

Implementation details. For EPIC-Kitchens-100 we use the TBN [31] features from the dataset provider comprising of 25 uniformly sampled RGB, flow, and audio feature vectors for each clip. For YouCook2 we use ImageNet-pretrained ResNet-152 features from the VALUE benchmark [38]. For the three methods we use the open source codebases provided in the respective papers and follow their hyper-parameter setting. We release our code and models on GitHub to support reproducibility.

4.2 Relevance-based margin results

EPIC-Kitchens-100. To validate the effectiveness of the proposed relevance-based margin, we explore three methods (MME, JPoSE, and HGR as described in Sec. 3.3) on EPIC-Kitchens-100. In Tab. 1 we report nDCG and mAP values, averaged between text-to-video and video-to-text. In all three cases, we observe a large improvement in both metrics, showing that the relevance-based margin works on very different models. It also works well with both offline mining with randomly sampled triplets (for MME and JPoSE), and online mining with hard negatives (for HGR): by using the relevance-based margin, MME gains +1.1 nDCG and +0.7 mAP, JPoSE +2.7 nDCG and +1.8 mAP, and finally HGR obtains +18 nDCG and +9.6 mAP. Such a large improvement is possibly due to how the triplets are sampled: in JPoSE, the negatives do not share the verb class of the anchor, leading to a relevance lower than 0.5; but, there is not such a guarantee in HGR, since batches are formed randomly. Hence, by employing a relevance-based margin in HGR we automatically deal with situations in which the negatives have a considerable relevance and adapt the margin accordingly. Finally, in App. A we report the public leaderboard for the retrieval challenge, confirming the improvement we observe over current state-of-the-art methods.

YouCook2. In the previous experiment we used the class knowledge which accompanies the dataset. But, by computing synsets knowledge in a similar way to what is done in EPIC-Kitchens-100, the proposed relevance-based margin can still successfully help the training process. This setting poses two additional challenges: first of all, in EPIC-Kitchens-100 most of the captions follow a precise structure, i.e. they contain a verb and a noun, which is not the case when dealing with other datasets, where free-form descriptions are often adopted. This may make it more difficult for the PoS-tagger to correctly tag the words. Secondly, there may be words which are put in the wrong category by WordNet.

For this dataset, we use the same class knowledge used in EPIC-Kitchens-100, as it transfers well across both datasets since they share the cooking domain [53], and for words which do not appear in any class, a new singleton class is created.

In Tab. 2 we report the nDCG and mAP values obtained with MME, JPoSE, and HGR. From the table, one can see that even in this different setting the relevance-based margin is able to provide useful information to the model. For example, the addition of the proposed technique in HGR leads to a gain of +5.5 nDCG and +3.1 mAP when compared to the results obtained with a fixed margin.

4.3 Ablation studies

We perform the ablation studies on EPIC-Kitchens-100 using JPoSE. Varying the fixed margin. In Sec. 4.2 we show that the proposed relevance-based margin leads to improved nDCG and mAP on both EPIC-Kitchens-100 and YouCook2. But, what if one would only need to carefully tune the fixed margin to obtain similar results? To answer to this question, we focus on JPoSE and vary the fixed margin $\Delta$ in $\{0.1, 0.2, \ldots, 1.5\}$ (default value used in JPoSE is 1).
| Method   | rel-Δ | nDCG | mAP  |
|----------|-------|------|------|
| MME      | ✓     | 46.9 | 19.3 |
| JPoSE    | ✓     | 47.3 | 19.5 |
| HGR      | ✓     | 41.0 | 23.0 |

Table 2: nDCG and mAP using MME, JPoSE, and HGR on YouCook2. We use ResNet-152 (pretrained on ImageNet) features from the VALUE benchmark [38].

| rel-Δ | cross- within- glob PoS glob+PoS | nDCG | mAP  |
|-------|--------------------------|------|------|
| ✓   | ✓                        | 53.1 | 43.3 |
| ✓   | ✓                        | 54.7 | 44.5 |
| ✓   | ✓                        | 53.4 | 43.7 |
| ✓   | ✓                        | 56.2 | 45.6 |
| ✓   | ✓ ✓                      | 56.2 | 45.8 |

Table 3: nDCG and mAP using JPoSE on EPIC-Kitchens-100. During training, JPoSE considers both cross- and within-modality contrastive losses, both at sentence- and PoS-level. Applying the relevance-based margin helps at each level.

We keep the rest of the hyper-parameter setting as in [15, 54] and use the officially provided TBN features. We plot in Fig. 2 nDCG, mAP, average R@1 for each of the tested margins. While small margins lead to worse results overall, it can be seen that increasing the margin does not improve significantly neither the nDCG nor the mAP. Moreover, the recall rates are affected only marginally as well. When compared to the performance shown by the adoption of the relevance-based margin, it can be observed that our technique manages to achieve higher nDCG and mAP values, while also keeping similar recall rates (on average, 6.3% R@1). Finally, it is worth noticing that by using the relevance-based margin we are released from the margin hyper-parameter: this is also of practical importance, because by using a fixed margin its optimal value is not known in a testing scenario, hence one would also need to perform an expensive search on the validation set in order to achieve better performance.

**Losses ablation.** A peculiarity of JPoSE is that it uses multiple contrastive loss terms to learn both global- and PoS-restricted joint embedding spaces. To do so, the authors employ a global loss and a PoS-level loss, both in the cross- and within-modality settings. We perform an ablation study in Tab. 3 to give evidence that the relevance-based margin can be helpful even when restricting the amount of loss terms used. Note that when applying the technique to the PoS-level terms (e.g. verbs) we consider the term for the opposite PoS (e.g. nouns) in Eq. 1 to be 1. As shown in Tab. 3, the adoption of the relevance-based margin leads to an improvement of +1.6 nDCG and +1.2 mAP when using only the cross-modal global-level loss terms, whereas +2.8 nDCG and +1.9 mAP are gained when also adding the cross-modal PoS-level terms.

**Modalities ablation.** For EPIC-Kitchens-100 we have RGB, flow, and audio features. To show that the improvements obtained when applying the relevance-based margin are not due to the model accessing multiple modalities related to the video, we perform another ablation in Tab. 4 by considering RGB-only and RGB+flow features. In both cases the proposed technique shows its usefulness. In particular, by employing the relevance-based margin we observe +1.6 nDCG and +1.6 mAP when using RGB-only, +2.9 nDCG and +1.8 mAP when using both RGB and flow, and +2.7 nDCG and +1.8 mAP when adopting all the three modalities.

**4.4 Qualitative analysis**

First of all, the proposed technique leads to variable margins, therefore the distribution of the values may help explaining why we observe such a positive influence on the final performance. In Fig. 3 we plot the frequencies of the margins (with bins of size 0.1) observed during the training of JPoSE on YouCook2, where for each of the training examples 10 triplets are sampled. It can be seen that a great part of the margins used are in the final bin (between 0.9 and 1.0), for which the relevance is quite low since the margin is...
Table 4: TBN offers RGB, flow, and audio features. The proposed relevance-based margin interacts with each modality in an incremental way. We use JPoSE on EPIC-Kitchens-100.

| Modalities | rel-Δ | nDCG  | mAP  |
|------------|-------|-------|------|
| RGB        | ✓     | 36.8  | 28.8 |
|            |       | 38.4  | 11.6 |
| RGB+Flow   | ✓     | 49.6  | 41.0 |
|            |       | 52.5  | 12.9 |
| RGB+Flow+Audio | ✓   | 53.5  | 44.0 |
|            |       | 56.2  | 15.8 |

Figure 3: Log-scale distribution of the margins used during training. Over each bin we report the frequency. Numbers refer to one epoch, with 10 triplets sampled for each example (e.g. with 10337 examples for YouCook2, we end up with around 103k triplets). Although a great part of the triplets are separated with the highest margin (i.e. lowest relevance), around 20k triplets are distanced by various margin values. Computed as \( \Delta_{a,p,n} = 1 - R(a,n) \) (see Eq. 7). In these cases, the margin will be similar to the default case of JPoSE, i.e. 1.0. Yet, around 20% of the training triplets end up having smaller margins. In these situations, the varying margins help the model achieve better performance by providing a semantic supervision on the structure of the embedding space, since the relevant items are kept at a distance which is proportional to the relevance.

Secondly, in Fig. 4 we visualize a few video-to-text examples from the testing set, by plotting for each of them the relevance values of each caption in both the full ranking list and the top 50 retrieved captions. By plotting the full ranking list, it is possible to see that the relevance-based margin helps improving the nDCG, as relevant captions are retrieved first. This can also be seen in the top 50 of Fig 4.a, 4.b, and 4.c where with the relevance-based margin no irrelevant captions are retrieved and, especially in Fig. 4.c, the ranking is almost ideal. Yet we can still find examples where the proposed technique fails to achieve the expected improvements. In Fig. 4.d, using the relevance-based margin a few irrelevant captions are retrieved, such as ‘take container’ and ‘take milk container’. This behavior is likely related to the fact that during training captions like ‘close container’ and ‘close milk container’ are relevant (0.5) for a video depicting the action ‘close fridge’, since they share the same verb class. This leads to an increase in the similarity of the respective descriptors. Hence, during inference, also captions like ‘take container’ and ‘take milk container’ might have a significant similarity with the video descriptor of ‘close fridge’. Further qualitative analysis is available in Appendix B.

5 CONCLUSIONS

Learning a joint embedding space using a margin-based contrastive loss is the dominant approach to deal with text-video retrieval. In the literature it is shown that by using such a framework, competitive performance on rank-unaware metrics, such as the recall rates, can be obtained. Yet, rank-aware metrics, such as the nDCG, need to be taken into account, as multiple descriptions can have numerous levels of relevance to a given query [53]. In this work, we proposed a technique to move away from the fixed margin which is typically used in such a framework, and introduced a relevance-based margin. In particular, we adopted the proposed technique into three increasingly more complex models on two datasets and gave empirical evidence that we can easily improve the performance measured through nDCG and mAP. Moreover, we showed that even by performing an expensive search of the fixed margin hyper-parameter, it does not reach the same performance as when using the relevance-based margin. Furthermore, the proposed technique can also have a positive impact on video retrieval applications as not needing to tune the margin can lead to less GPU hours required to fully train the model. Finally, we focused our work on text-video retrieval, but the relevance-based margin can be easily extended to other domains where similar margin-based ranking losses are used, e.g. in image retrieval [58]. Moreover, we showed the effectiveness of the proposed approach by applying it to loss functions where the margin is explicitly defined and used to separate positive and negative pairs, e.g. [9, 48]. Yet, there are also popular loss functions which do not make use of it, such as NCE [23] and MIL-NCE [43]. Future work is required to adapt the relevance-based margin to non-margin based loss functions.

ACKNOWLEDGMENTS

We gratefully acknowledge the support from Amazon AWS Machine Learning Research Awards (MLRA) and NVIDIA AI Technology Centre (NVAITC), EMEA. We acknowledge the CINECA award under the ISCARA initiative, which provided computing resources for this work. This work has been partially supported by the Spanish project PID2019-105093GB-I00 and by ICREA under the ICREA Academia programme.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition. 6077–6086.

[2] Stanisław Antol, Ashwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision. 2425–2433.
Figure 4: Video-to-text qualitative results on EPIC-Kitchens-100 testing set using JPoSE. For each of the examples we show a few frames and the groundtruth (GT) caption, and we plot both the full ranking list and the top 50 retrieved captions when adopting the fixed margin and then the relevance-based margin. On the left we also visualize the color bar which is used for the relevance (light colors mean high relevance, dark colors low relevance). In particular, Figures a, b, and c are success cases, whereas Figure d represents a failure case.

[3] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. 1999. Modern information retrieval. Vol. 463. ACM press New York.

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. ICCV (2021).

[5] Mathieu Blondel, Olivier Teboul, Quentin Berthet, and Josip Djolonga. 2020. Fast differentiable sorting and ranking. In International Conference on Machine Learning. PMLR, 950–959.

[6] Christopher Burges, Robert Ragno, and Quoc Le. 2006. Learning to rank with nonsmooth cost functions. Advances in neural information processing systems 19 (2006), 193–200.

[7] Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent.
We show the results of both the participants at the time of submission and those that took part into the previous challenge (which ended in November 2021) in Figure 5. The previous best result was obtained by Hao et al. (more details in the technical report [16]), which achieved on average 44.23% mAP and 53.56% nDCG. As can be seen, we achieve 45.86% mAP (+1.63%) and 56.21% nDCG (+2.65%).

B Qualitative Analysis

We further analyze the effectiveness of the proposed technique from a qualitative point of view. To do so, we select three types of information. First of all, we pick a caption and compute its embedding ($q$), pick the corresponding video descriptor ($v$), and compute their similarity $s(v, q)$ through dot product. Then, we look for 10 similar captions (i.e. different captions which either share the noun or the verb class), pick the corresponding video descriptors indexed by $V^+$, and compute an aggregated similarity value $s(v^+, q) = \frac{1}{10} \sum_{v_i \in V^+} s(v_i, q)$. Finally, we also randomly select 10 dissimilar captions (i.e. sharing neither the verb nor the first noun class), pick their video descriptors, and compute $s(v^-, q)$. We compare the results using JPoSE on the testing set of EPIC-Kitchens-100, and report several examples in Figure 6. In Figures 6.a and 6.b the usage of a fixed margin leads to a far too high similarity of the videos in $V^+$ with the query $q$ when compared to its groundtruth video descriptor $v$, which hurts both nDCG, mAP, and the recall rates. In Figures 6.c and 6.d the videos in $V^-$ and those in $V^+$ are not properly separated, hence wrongly giving the model the impression that they are similarly relevant to the query $q$. In all these cases, adopting a relevance-based margin is a successful strategy to correct these wrong predictions, leading to a more robust model.

A Comparison with the EPIC-Kitchens-100 Challenge Leaderboard

The release of the EPIC-Kitchens-100 dataset [15] was accompanied by a public challenge for the multi-instance retrieval problem (alongside other challenges, e.g. for Action Recognition). To further prove the results we show in Section 4, we took part into the challenge by employing the proposed relevance-based margin on the JPoSE method [54] (see Section 3).
Figure 5: We report the public leaderboard for the EPIC-Kitchens-100 Challenge at time of submission (below), and also the leaderboard for the previous challenge which ended in November 2021 (above). It can be seen that we achieve around +1.6% mAP and +2.6% nDCG over the previous best results, achieved by Hao et al. (details in the technical report [16]).
Figure 6: Using JPoSE, we compute a similarity score $s(\alpha, q)$ for the groundtruth pair (colored in blue), $s(\alpha+q)$ for videos with similar captions (colored in green), and $s(\alpha-q)$ for videos with dissimilar captions (colored in brown). Note that, when selecting $V_s$ for the examples on the left we change the noun class, whereas on the right we change the verb class. See Sec. B for more details. The captions of the videos used are reported on the right. Each of the four examples are taken from EPIC-Kitchens-100 testing set and for each of them we report first what happens with fixed margin, then with the proposed relevance-based margin.