Learning Joint Embeddings for Video-Text Retrieval for Higher Coverage

Jayaprakash A*
Abhishek*
akulajayaprakash@gmail.com
abhishek.at3@gmail.com
Indian Institute of Technology, Bombay
Mumbai, Maharashtra, India

Rishabh Dabral
Indian Institute of Technology, Bombay
Mumbai, India
larst@affiliation.org

Vighnesh Reddy
Indian Institute of Technology, Bombay
Mumbai, India

Preethi Jyothi
Indian Institute of Technology, Bombay
Mumbai, Maharashtra, India

Ganesh Ramakrishnan
Indian Institute of Technology, Bombay
Mumbai, Maharashtra, India

ABSTRACT

Video retrieval using natural language queries requires learning semantically meaningful joint embeddings between the text and the audio-visual input. Often, such joint embeddings are learnt using pairwise (or triplet) contrastive loss objectives which cannot give enough attention to 'difficult-to-retrieve' samples during training. This problem is especially pronounced in data-scarce settings where the data is relatively small (10% of large scale MSR-VTT) to cover the rather complex audio-visual embedding space. In this paper, we propose to compensate for data scarcity by using domain-knowledge to augment supervision. To this end, in addition to the conventional three samples of a triplet (anchor, positive, and negative), we introduce a fourth term - a *partial* - to define a differential margin based partial-order loss. The *partials* are heuristically sampled such that they semantically lie in the overlap zone between the positives and the negatives, thereby resulting in broader embedding coverage. Our proposals consistently outperform the conventional max-margin and triplet losses and improve the state-of-the-art on MSR-VTT and DiDeMO datasets. We also release a multilingual video-text retrieval dataset which has relatively fewer number of examples in each language in comparison to existing datasets. We report benchmark results while also observing significant gains using the proposed partial order loss, especially when the language specific retrieval models are jointly trained by availing the cross-lingual alignment across the language-specific datasets.

KEYWORDS

datasets, neural networks, gaze detection, text tagging

*Both authors contributed equally to this research.

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

ACM Reference Format:
Jayaprakash A, Abhishek, Rishabh Dabral, Vighnesh Reddy, Preethi Jyothi, and Ganesh Ramakrishnan. 2018. Learning Joint Embeddings for Video-Text Retrieval for Higher Coverage. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Learning low-dimensional, semantically rich representations of audio, visual and textual data (associated with videos) is a fundamental problem in multimodal learning [15, 16, 18]. Such multimodal representations play a critical role in facilitating a multitude of visual analytics tasks such as Video Retrieval [17], Visual Question Answering (VQA), video summarization *etc.*

Learning such compact representations is already a challenging task owing to the high dimensionality and multi-modality of sensory data. This difficulty is further exacerbated when the available data is limited and biased. In this paper, we propose a novel approach to learning joint-embeddings between videos (audio-visual) and text while specifically being cognizant of its effectiveness for data-scarce settings.

A typical approach to learning joint video-text embeddings involves projecting the video and text inputs onto a common latent space such that semantically similar video-text pairs are in the same neighbourhood whereas dissimilar video-text pairs are pushed far apart. This projection is achieved by training video and text encoders using triplet ranking losses which operate on triplets of <anchor, positive, negative> samples [3]. Here, the anchor can be a video embedding and the positive and negative parts can correspond to instances with textual embeddings or vice-versa.

While this training regime is suitable for resource-rich languages with availability of large audio-visual and textual datasets, the performance quickly deteriorates when applied to data-scarce settings due to two main reasons. Firstly, the scarcity of data weakens the video (and textual) encoders which do not *cover* the shared embedding space densely enough, thereby restricting their ability to meaningfully project to the shared space. Secondly, data

scarcity inevitably prevents the encoders from generalizing to out-of-distribution text or video queries. A natural solution to the problem would be to explicitly target the hard positive and hard negative instances in the data. This approach, however, has limited returns as it leads to noisy gradients, thereby leading to a collapsed model [25].

In this paper, we propose an alternative solution in the form of a differential margin based *partial order contrastive loss* that facilitates augmenting the supervision using domain knowledge. We argue that the triplet formulation, while crucial, is not sufficient. To further hand-hold the training process, we propose to learn the embeddings using a quadruplet of <anchor, positive, *partial*, negative> instances. The additional *partial* samples are sampled using meaningful heuristics from domain knowledge such that their semantic interpretations fall somewhere in between the positive and negative samples. We demonstrate that these heuristics can be almost as effective as manually identified *partial* samples that we also release as part of the dataset (RUDDER) associated with this paper.

We illustrate the intuitiveness of our proposal with the help of an example in Figure 1, showing a frame from a video of a man watching television. While the positive and negative sentences, *'A person sitting on a sofa watching television'* and *'A woman is throwing a dress in the air'* are self-explanatory, the partially relevant sentence, *'A person is watching videos on a laptop'* is neither a strictly positive nor a strictly negative instance. Though the objects are different, the sentences still capture the act of a person watching a (similar-looking) object. We postulate that placing such sentences at an appropriate distance from the anchor can crucially contribute towards coverage within the sparse latent space shared between the videos and the text. When training with partial instances, we differentiate them from the anchors using a partial-margin value that lies in between the margin values of positive and negative samples. Note that the proposed setup is different from the similar-sounding paradigm of semi-hard negative mining [19]. While semi-hard negative mining is a useful tool to ease the training process, it does not deal with the issue of coverage over the embedding space.

We show that our partial-order loss boosts the projective power of encoders, especially for data-scarce settings, and consistently outperforms triplet-based ranking losses. We also show that while specifically designed for data-scarce settings, the method extends to larger datasets such as MSR-VTT and Charades, demonstrating considerable improvements in retrieval performance. Furthermore, these results are achieved with minimal hyperparameter tuning of the partial margins. To demonstrate the efficacy of our new loss function, we employ a state-of-the-art "Collaborative Experts" architecture [16] that offers multimodal shared embeddings learnt jointly by utilizing features from various modalities including audio, object, motion, scene and action.

We also attempt to address another limitation of existing objectives employed in the joint learning of embeddings. The commonly employed contrastive loss [8] or triplet loss [3] only consider the semantic information within individual pairs or triplets of examples, respectively, while ignoring the interactions with the remaining examples. Neither the objective nor the update steps give enough weightage to hard positives (matching sample that is farthest from each training query) or hard negatives (non-matching sample closest to each training query) based on their hardness. We consider

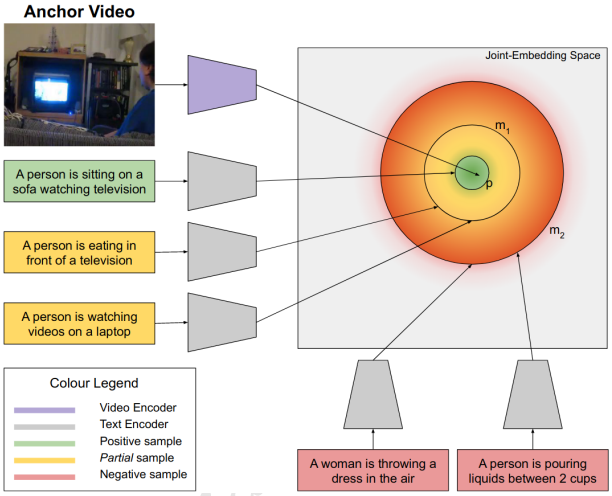

Figure 1: Figure illustrating the different roles played by positive, *p*artially overlapping and negative samples in our proposed simple differential margin based partial-order loss. Intuitively, we would like distances between the partially similar pairs to lie in the dark orange zone between the positives and the negatives, *i.e.*, beyond the m_1 margin but within m_2. Distances between unrelated pairs should be pushed to the grey region beyond m_2, whereas distances between positive pairs should be in the green region within the margin p.

an optimal transport-based loss that differentially weighs the hard examples in contrast against the easier ones. Such a loss could be further enhanced using augmented supervision.

In summary, the contributions of this paper are as follows:

1) A new quadruplet-based partial order loss is proposed to improve the coverage of video-text retrieval models. The proposed intuitive loss takes into account augmented supervision along with the already available ground truth relevance. In this loss, we tune margins that help differentiate positives from the partial and the partial from the negatives.

2) We release a newly annotated cross-lingual video dataset, RUDDER[1], to test our (and relevant) methods on data-scarce settings. The dataset includes audio-visual footages in Marathi, Hindi, English, Tamil and Telugu. A significant percentage (75%)of the videos are the same, and for these videos, the audio tracks are largely aligned across the languages. Further, for each of the aforementioned languages, for every video that has audio track in that language, we provide captions in that language along with the associated start and end timings.

2 RELATED WORK

Learning Representations for Video-Text Retrieval: Several of the existing approaches for video-text retrieval[9, 18, 24] are based on image-text embedding methods by design, and typically focus on single visual frames. [9] encode video features directly into the word2vec embedding space by using Hierarchical Recurrent Neural Network (HRNNs) and compare them with class labels

[1]RUDDER stands for c**R**oss ling**U**al vi**D**eo an**D** t**E**xt **R**etrieval.

projected into the same space. [24] focus on learning separate embedding spaces for each POS tag, such as nouns, verbs, or adjectives using triplet losses wherein the positive examples for the triplet loss were constructed by finding relevant entities that shared common nouns and verbs. We borrow inspiration from this work to design our heuristics. [17] propose Mixture of Embedding Experts to learn from heterogeneous data sources representing different concepts. [18] show the effective utilization of multimodal cues in learning better joint embeddings for cross-modal retrieval using pairwise loss functions with an emphasis on hard negatives. Their framework made use of a fusion of action, object, place, text and audio features extracted from state-of-the-art models. Likewise, Collaborative Experts [16] is proposed as a network to effectively fuse the aforementioned features (experts) with the reasoning that extracting supervision from multiple experts could help compensate for the lack of sufficient data. Our proposed loss function is used within the collaborative experts framework to further help coverage, specifically in low-resource settings.

Loss Functions for Cross-Modal Learning: Several max-margin based ranking loss functions have been proposed for learning joint embeddings. The triplet loss [3, 19] and bidirectional max-margin ranking loss [20] have been popularly used to learn cross-modal embeddings. [28] propose a new approach to learn distance metrics via optimal transport programming for batches of samples and report faster convergence rates without adversely affecting performance. Significant effort has also gone into better sampling methods for training ranking losses. [19] use semi-hard negatives within the triplet-loss framework. Notably, [25] propose a distance-based sampling method to choose stable and informative samples that can be trained with a simple max-margin loss. Likewise, [23] propose to consider samples in the neighborhood of positives. Our partial ordering based loss formulation subsumes it in the sense that we also consider the extended neighborhood between the positives and negative. The approach closest to our partial-order based loss, however, is [14] which uses the knowledge of categorical hierarchy to label samples as positives or semi-positives. In addition to differing from them in our loss formulation, it is worth noting that unlike categorical data, establishing hierarchies in textual queries is not straightforward.

3 OUR METHODOLOGY

We propose a quadruplet-based partial-order augmented contrastive loss to learn joint video-text embeddings. We also show how this augmented supervision can be used to improve an OT-based loss function [28].

Let $\mathcal{V} = \{v_1, v_2, \ldots, v_{|\mathcal{V}|}\}$ and $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$ denote a set of videos and captions. We define $f(\cdot)$ and $g(\cdot)$ as video and caption encoders, respectively, both of which embed an i^{th} video, v_i, and a j^{th} caption, t_j, into an N-dimensional space. Let $d_{i,j} = \text{dist}(f(v_i), g(t_j))$ be the distance between the embeddings corresponding to video v_i and caption t_j. The standard *bidirectional max-margin ranking loss* [20] can be written as:

$$\mathcal{L} = \sum_{i,j \neq i} [m + d_{i,i} - d_{i,j}]_+ + [m + d_{i,i} - d_{j,i}]_+, \quad (1)$$

Here, m is a tunable margin hyperparameter that lower-bounds the distance between negative pairs and $[.]_+$ represents the $\max(., 0)$ function.

3.1 Partial Order Contrastive Loss

The bidirectional max-margin loss defined above separates the negative instances from positive instances by a fixed margin, m. Although this is regarded as one of the standard ways of learning cross-modal embeddings, we argue that in the absence of a large dataset, the loss results in a sparsely *covered* embedding space, thereby hurting the generalizability of the embeddings. Furthermore, the pretrained embeddings (GLoVE, GPT, etc) for low-resource languages are often either weak or simply may not exist.

In order to circumvent the challenges associated with a low-resource language setting, we propose the following quadruplet based mining setup. Given a batch of videos and their corresponding captions, for every anchor sample, we construct three sets of video-caption pairs, *viz.*, (i) the positive (S^+), (ii) the negative (S^-) and (iii) the *partial*-overlap ($S^\sim$). While the positive and negative pairs are chosen as in the bidirectional max-margin loss, we show we can make effective use of dataset-dependent heuristics to sample the *partial* samples. Intuitively, the *partial* samples are chosen such that they represent a degree of semantic overlap with the anchor (as illustrated in Figure 1). We formally define the proposed novel Partial Order (PO) Contrastive Loss as:

$$\mathcal{L}^{PO} = \mathcal{L}^+ + \mathcal{L}^- + \mathcal{L}^\sim$$

$$\mathcal{L}^+ = \sum_{(i,j) \in S^+} [d_{i,j} - d_{i,i} - p]_+ + [d_{j,i} - d_{i,i} - p]_+$$

$$\mathcal{L}^- = \sum_{(i,j) \in S^-} [n + d_{i,i} - d_{i,j}]_+ + [n + d_{i,i} - d_{j,i}]_+$$

$$\mathcal{L}^\sim = \sum_{(i,j) \in S^\sim} \Big([m_1 + d_{i,i} - d_{i,j}]_+ + [m_1 + d_{i,i} - d_{j,i}]_+ + [d_{i,j} - d_{i,i} - m_2]_+ + [d_{j,i} - d_{i,i} - m_2]_+ \Big)$$

Here, p, m_1, m_2 and n are tunable margin hyperparameters such that $p < m_1 < m_2 < n$. As depicted in Figure 1, we would like distances between the partially similar pairs to lie in the dark orange zone between the positives and the negatives, *i.e.*, beyond the m_1 margin but within m_2. Distances between unrelated pairs should be pushed to the blue region beyond m_2, whereas distances between positive pairs should be in the green region within the margin p. Heuristics for selecting partially overlapping samples is described next.‘

3.2 Choosing the Heuristics

The success of the proposed partial order loss function relies on an appropriate choice of the heuristics. Since the heuristics depend on the complexity of captions and the content of videos, we describe a generic heuiristic in Figure **??**. In most datasets (Charades-STA, RUDDER, DiDeMO, etc), the textual captions consist of a subject, one or more objects (noun) and the associated verbs, all of which can be retrieved using an off-the-shelf part-of-speech POS tagger. Given the recovered nouns and verbs, we compute the percentage overlap between the noun-verbs of the query caption with all the other captions in the batch. Trivially, a 100% overlap represents a positive sample and a 0% overlap represents a negative sample.

For the partial sample selection, however, we select thresholds α_n and α_v for the percentage overlap of nouns and verbs, respectively. Finally, a sample is labeled as partial if it is *not* a positive sample and satisfies either the noun threshold or the verb threshold. Intuitively, a common verb is a strong signal of similarity between two actions, hinting towards the possibility of similar visual features. Likewise, the presence of common objects in two videos also restricts the space of possible visual and textual features. The values of α_n and α_v depend on the dataset and are empirically determined using a hyperparameter sweep. Note that this is only one of many possible heuristics and we empirically demonstrate that this simple approach is nearly as effective as the one in which partially relevant samples are manually identified. Specifically, for the RUDDER dataset, we manually labeled the captions using a description-based heuristic (discussed later) which yields comparable performance (*c.f.* Table 1).

4 EXPERIMENTAL SETUP

We will first describe the datasets in detail, followed by implementation details and our experimental results.

The RUDDER Dataset: We release a cross-lingual video and text retrieval dataset to help evaluate the efficacy of our proposal on a set of low-resource languages. The dataset consists of video tutorials for making scientific toys from waste material with audio narrations available in multiple Indian languages such as Marathi, Telugu, Tamil and Hindi[2].Captions were manually annotated by native speakers in case of Marathi and Hindi. We have 3272 videos overall, of which we use 2287 for training, 166 for validation and 819 for testing. All these videos have audio in Marathi and additionally most of these have audio additionally in other languages. In the case of Telugu and Tamil, we use the corresponding Google Translate APIs for obtaining the captions. With annotations in a low-resource language (Marathi), we use RUDDER to showcase the utility of our partial order loss. We apply two heuristics to the RUDDER dataset to generate the partially relevant instances: a) Noun-Verb heuristic (as discussed earlier in Section 3.2), and b) a description-based manually annotated heuristic. For the description-based heuristic, we manually partition the captions into two categories: captions *listing* the objects - *eg.*, you need plastic bottles, cycle spoke, sticks, ring magnet and steel pin. (English translation) - and the captions *describing* how to use the listed objects - *eg.*, place two strong ring magnets in cycle spoke. Furthermore, we ask the annotators to manually identify nouns and verbs in the sentence. These annotations, which we will publicly release, are used to extract the partial samples. The reader is referred to Supplementary Material for more details.

Of the 3272 videos, 675 (roughly 21%) have audio available in all languages. On the other hand, 54% videos are in both Hindi and Marathi, 27% in Tamil and Marathi and 23% in Telugu and Marathi. In each setting (including cross lingual), we use 70% as train data, 1% as validation data and 24% as test data.

Charades-STA:. Charades-STA [7] is a human-centric activity dataset with sentence-level temporal annotations for each video. The dataset is challenging for video retrieval tasks as it contains only one, often short, caption per video. We use 4260 clips in total with 3124, 121 and 1015 for training, validation, and testing, respectively, while ensuring that each caption is uniquely mapped to one video.

MSR-VTT:. MSR-VTT [27] is a relatively large video dataset consisting of 10k videos with English captions. We follow a train-val-test split of 6512, 496, and 2990, respectively. Unlike Charades-STA, it provides multiple English captions per video. Similar to [16], we conduct experiments under two settings: 1) with 20 captions per video, and 2) with only 1 caption per video. We hypothesize that in the resource constrained setting of 1 caption per video, our method should outperform the existing methods.

DiDeMo: Distinct Describable Moments (DiDeMo) [10] dataset consists of over 10k videos in diverse visual settings with pairs of localized video segments and referring expressions. Each video has around 3-5 descriptions aligned with time-stamps. We fuse the captions into a single description. We follow a train-val-test split of 8392, 1065 and 1004 videos respectively.

4.1 Evaluation Metrics

We use standard retrieval measures adopted from prior work [6, 15, 17] and show results on both text-to-video (T2V) and video-to-text (V2T) retrieval. R@K (recall at rank K) for K=1,5,10,50 measures the percentage of the top-K retrieved text/video results, for a given video/text query, that match the ground-truth. Median Rank (MdR, lower is better), and Mean Rank (MnR, lower is better) are the two other measures we consider, that compute the median and mean of the ground truth appearing in the ranking of the predictions, respectively. When computing video to-sentence measures for datasets containing multiple independent sentences per video, such as the RUDDER dataset, we follow the evaluation protocol used in prior work [4, 5] which corresponds to reporting the minimum rank among all valid text descriptions for a given video query. To determine convergence, we use two parameters maximum geometric mean for R@5, R@10, & R@50 and minimum geometric mean of MnR and MdR.

4.2 Implementation Details

Video Encoder: We adopt the state-of-the art collaborative experts (CE) framework from [16] as our base model. We chose four pretrained models to serve as experts and derived features from the scene [2, 13], audio (VGG [11]), objects [12, 26] and action [21], that are subsequently fed into the video encoder as inputs. The embeddings from these features are further aggregated within the video encoder using a collaborative gating mechanism, resulting in a fixed-width video embedding (previously denoted as $f(v)$).
Textual Encoder: In order to construct textual embeddings, denoted by $g(t)$, the query sentences are converted into a sequence of feature vectors using pretrained word-level embeddings like GloVE and OpenAI-GPT. These word level embeddings are aggregated using NetVLAD [1]. We subject the output of the aggregation step to the text encoding architecture proposed by [17], which projects text features to different subspaces, one for each expert. Each text feature is further refined by computing attention over aggregated text features and thereafter passed through a feedforward and a

[2]We downloaded these videos from http://www.arvindguptatoys.com/toys-from-trash.php and obtained consent from the content creator. We will release the dataset on acceptance of our paper.

Text to Video Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
Triplet	1.91	7.45	12.13	32.80	108.33	178.33
DW	2.16	7.45	12.13	33.41	110.33	184.11
MM	2.16	8.63	14.29	36.39	99.67	180.15
OT	2.85	9.40	14.86	**37.53**	99.33	**175.23**
PO	**3.01**	**10.70**	**15.91**	36.39	**97.67**	181.71
PO(M)	**4.48**	**13.47**	**20.02**	**44.28**	**66.00**	**153.14**

Video to Text Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
Triplet	1.99	7.33	12.09	33.66	107.00	173.32
DW	1.99	7.20	11.40	31.83	113.33	184.22
MM	2.12	9.08	13.84	36.18	96.00	174.87
OT	2.85	9.48	14.94	37.61	94.00	**172.77**
PO	**3.13**	**10.50**	**16.32**	**38.34**	**93.33**	180.06
PO(M)	**3.87**	**12.13**	**19.09**	**42.49**	**73.00**	**151.63**

Table 1: Results on human annotated RUDDER dataset. PO and PO(M) denote the use of heuristically annotated and manually annotated partial samples for the partial order loss. DW refers to our implementation of Distance-Weighted sampling as in [25].

softmax layer to produce a scalar for each expert projection. Finally, each expert is scaled and concatenated to form a fixed-width text embedding.

Training: We use PyTorch for our implementation. The margin values p, m_1, m_2 and n are tuned on the validation sets of each dataset. More details are provided in the supplementary material.

4.3 Baselines

We contrast and compare our improvisation of the collaborative experts (CE) model trained on our simple differential margin based partial order loss against the following baselines (i) CE trained on the vanilla max-margin loss (MM) in Equation (1) (ii) an importance-driven distance metric via batch-wise Optimal Transport (OT) (see supplementary for details). (iii) a Triplet loss (Triplet) baseline, (iv) HN baseline which refers to the hard negative mining strategy, where we pick only the hardest negative to train on and, finally, (v) a distance-weighted sampling based triplet loss (DW) from [25], (vi) S2VT (Sequence to Sequence Video Text retrieval), which is a reimplementation of [22], (vii) FSE [29], which performs Cross-Modal and Hierarchical Modeling of Video and Text and (viii) specifically on the MSRVTT and DiDeMo datasets, we also present the numbers as reported by [16] and refer to those as (MM[Liu]). To the best of our knowledge, we are the first to evaluate the effectiveness of OT based loss for retrieval tasks.

5 RESULTS AND ANALYSIS

Results on RUDDER. Table 1 presents our main results on the RUDDER dataset using manually-derived annotations and heuristically-derived annotations. Further, in Table 2, we present the results on the multi-lingual RUDDER setting wherein the textual captions are in Marathi but audio-experts corresponding to other languages

Text to Video Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
MM	**4.73**	15.02	24.69	**62.96**	33.00	47.15
OT	3.5	16.67	27.98	62.14	31.67	45.8
PO	4.12	**17.49**	**29.42**	61.93	**31.00**	**45.74**

Video to Text Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
MM	5.35	**18.11**	25.31	62.55	32.83	46.65
OT	5.76	16.05	26.75	67.49	29.17	**41.95**
PO	**5.97**	16.46	**26.95**	**68.93**	**27.83**	42.09

Table 2: Results on the RUDDER dataset that is (multilingually) trained using audio from all the 4 languages. Multi-lingual RUDDER uses audio experts from four different languages (Hindi, Marathi, Tamil, Telugu).

are also used. For both these losses, we sample a single negative example from each batch which partly explains the deterioration in performance compared to MM. More details about these implementations are available in the supplementary material. We also compare against an OT-based loss function (OT).

We observe that adding augmented supervision and invoking our partial order loss yields consistent improvements across all the evaluation metrics, in both Tables 1 and 2. We also observe that multi-lingual training significantly improves the retrieval performance. These numbers clearly validate our claim about the utility of the partial order loss in limited-data conditions. To further confirm the effectiveness of the method, we perform similar experiments with textual queries from languages other than Marathi, such as Tamil, Telugu and Hindi (a mix of unrelated languages). Table 3 tabulates performances when both the audio and the text corresponds to the same language, whereas Table 4 tabulates results when the text is in Marathi but the audio features are derived from a different language. Additionally, we observe that our results with heuristically annotated RUDDER are fairly comparable to those obtained with the manually annotated RUDDER, thus proving that one can achieve good performance with reasonable heuristics without having to undertake any additional human evaluations.

	T2V Retrieval		V2T Retrieval	
Loss	MdR	MnR	MdR	MnR
MM(hin)	68.17	108.98	67.67	**107.32**
PO(hin)	**63.5**	**108.49**	**59.67**	110.19
MM(tam)	38.83	56.09	33.5	**50.44**
PO(tam)	**28.5**	**53.74**	**32.83**	50.57
MM(tel)	56.	85.25	53.17	81.99
PO(tel)	**52.83**	**83.99**	**51.67**	**81.58**

Table 3: Results on RUDDER dataset with audio and captions in the same language.

Additional Results: We also demonstrate the performance of the proposed loss function on three benchmark datasets: MSR-VTT, DiDeMO and Charades-STA. MSR-VTT is a large-scale dataset with up to 20 captions per video. To conform better to our low-resource

	T2V Retrieval		V2T Retrieval	
Loss	MdR	MnR	MdR	MnR
MM(hin)	125.17	143.19	124.	142.99
PO(hin)	**119.67**	**141.49**	**122.33**	**142.93**
MM(tam)	54.67	56.41	45.5	**52.94**
PO(tam)	**46.67**	**56.38**	**43.67**	54.07
MM(tel)	50.5	84.21	50.25	83.73
PO(tel)	**35.00**	**68.09**	**40.17**	**70.81**

Table 4: Results on RUDDER dataset with audio in different languages and captions in Marathi.

Text to Video Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
MM[Liu]	4.80	16.20	25.00	-	43.30	183.10
MM	5.61	18.18	27.62	56.91	35.33	**153.75**
OT	5.60	18.14	27.34	55.42	38.00	176.24
PO	**6.19**	**19.23**	**28.75**	**57.62**	**33.33**	161.29

Video to Text Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
MM[Liu]	8.40	25.60	37.10	-	20.30	87.20
MM	10.55	29.44	41.49	72.90	15.92	**69.53**
OT	9.39	27.66	39.61	70.19	17.67	83.87
PO	**11.22**	**30.71**	**43.40**	**73.12**	**14.67**	74.07

Table 5: Results on MSRVTT dataset using 1 caption.

setting, we use only 1 caption per video[3]. Table 5 shows the results with our loss functions, along with the numbers reported in [16] for this setting (MM[Liu]). We observe that PO outperforms all other loss functions on all metrics (with the exception of MnR), and significantly improves over MM[Liu]. Tables 6 and 7 shows results on the DiDeMO and Charades-STA datasets, respectively. PO performs fairly well on the Charades-STA dataset. On DiDeMO, it improves upon the state-of-the-art on MdR and higher *sk* values for R@k. Interestingly, we observe a deviation from the trend for R@1 aznd R@5 for the V2T task. We believe this can be attributed to the captions in DiDeMO being of fairly large length on average, which hampers the ability of simple heuristics such as our noun-verb heuristic to identify useful partial samples.

Significance-test: The significance of gains of PO over baselines such as Triplet, DW, CE and HN are evident from the Tables 1, 7 and 6. Hence we report Wilcoxon's signed-rank test on the median ranks of PO and the tightly contending MM losses for all the experiments in the paper. We observe statistical significance at a p-value less than 0.0001 in favour of PO against MM. We also observe a p-value less than 0.01 for PO against OT.

6 CONCLUSION

We present an approach to learning representations in a multi-modal retrieval setting that helps significantly improve recall while being sufficiently competitive with respect to precision. We propose a new partial ordering-based contrastive loss that makes use of (partially ordered) augmented supervision for training the embeddings with the provision that such additional augmented supervision could be generated either manually or automatically using simple heuristics. Through the augmented supervision, our approach learns to focus on 'difficult-to-retrieve' samples during optimization unlike recent work that focuses either on semi-hard-negative instances (to boost performance) or distance-weighted sampling. Our partial-order loss yields significant performance gains (wrt standard performance measures) over these approaches, as well as over standard pair-wise (or triplet) based loss objectives such as max-margin. We also validate the feasibility of heuristically creating augmented supervision and we release a newly annotated multilingual dataset. We also evaluate the positive effect of such augmented supervision on several other standard benchmarks in video-text retrieval.

[3]We also show results on the complete MSR-VTT dataset using all captions in the supplementary material. We do not see benefits from PO over the other losses in this setting given that evaluation using 20 captions has already a somewhat saturated coverage.

Text to Video Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
S2VT	11.9	33.6	-	76.5	13	-
FSE	13.9	36.0	-	78.9	11	-
MM[Liu]	16.1	41.1	-	82.7	8.3	43.7
HN	14.97	37.55	51.46	80.31	10.	47.62
MM	**16.43**	**42.03**	54.48	82.17	**8.00**	44.36
OT	15.03	41.73	54.98	82.66	9.00	41.23
PO	16.33	41.43	**56.47**	**84.46**	**8.00**	**40.21**

Video to Text Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
S2VT	13.2	33.6	-	76.5	15	-
FSE	13.1	33.9	-	78.0	12	-
MM[Liu]	15.6	**40.9**	-	82.2	8.2	42.4
HN	14.11	37.18	51.13	79.22	10.	42.42
MM	**16.93**	39.94	52.29	81.77	10.00	42.62
OT	14.14	38.84	53.48	82.47	9.00	**38.31**
PO	14.94	39.84	**54.88**	**84.56**	**8.00**	39.60

Table 6: Results on DiDeMo dataset

Text to Video Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
Triplet	1.59	5.57	9.18	26.18	153.00	240.48
DW	1.69	5.80	9.68	27.40	147.00	236.06
MM	2.43	8.87	14.34	39.71	89.00	174.57
OT	2.73	**9.95**	15.45	40.32	80.50	171.43
PO	**3.64**	9.72	**15.89**	**40.38**	**77.00**	**162.34**

Video to Text Retrieval

Loss	R@1	R@5	R@10	R@50	MdR	MnR
Triplet	1.32	4.22	8.03	25.00	157.17	251.97
DW	1.08	4.76	7.83	25.78	150.25	240.23
MM	1.99	8.40	13.36	38.19	86.83	174.43
OT	1.96	8.60	14.24	38.36	**82.33**	172.32
PO	**3.24**	**9.11**	**14.88**	**39.57**	83.00	**164.57**

Table 7: Results on Charades-STA dataset.

REFERENCES

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*.

[2] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*.

[3] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio. 2010. Large Scale Online Learning of Image Similarity Through Ranking. *Journal of Machine Learning Research* (2010).

[4] J. Dong, X. Li, and C. G. M. Snoek. 2018. Predicting Visual Features From Text for Image and Video Caption Retrieval. *IEEE Transactions on Multimedia* 20, 12 (2018), 3377–3388.

[5] Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. 2018. Dual Encoding for Zero-Example Video Retrieval. *arXiv e-prints*, Article arXiv:1809.06181 (Sept. 2018), arXiv:1809.06181 pages. arXiv:1809.06181 [cs.CV]

[6] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc' Aurelio Ranzato, and Tomas Mikolov. 2013. DeViSE: A Deep Visual-Semantic Embedding Model. In *NeurIPS*.

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *ICCV*.

[8] Raia Hadsell, Sumit Chopra, and Yann LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*.

[9] Meera Hahn, Andrew Silva, and James M. Rehg. 2019. Action2Vec: A Crossmodal Embedding Approach to Action Learning. In *CVPR-W*.

[10] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2018. Localizing Moments in Video with Temporal Language. In *EMNLP*.

[11] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. 2017. CNN architectures for large-scale audio classification. In *ICASSP*.

[12] Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2020. Squeeze-and-Excitation Networks. *IEEE TPAMI* (2020).

[13] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227* (2014).

[14] K. Karaman, E. Gundogdu, A. Koç, and A. A. Alatan. 2019. Quadruplet Selection Methods for Deep Embedding Learning. In *ICIP*.

[15] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *arXiv e-prints*, Article arXiv:1411.2539 (Nov. 2014), arXiv:1411.2539 pages. arXiv:1411.2539 [cs.LG]

[16] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. 2019. Use What You Have: Video retrieval using representations from collaborative experts. In *BMVC*.

[17] Antoine Miech, Ivan Laptev, and Josef Sivic. 2019. Learning a Text-Video Embedding from Incomplete and Heterogeneous Data. In *ICCV*.

[18] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. 2018. Joint embeddings with multimodal cues for video-text retrieval. *International Journal of Multimedia Information Retrieval* 8 (2018), 3–18.

[19] F. Schroff, D. Kalenichenko, and J. Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*.

[20] Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* (2014).

[21] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *CVPR*. 6450–6459.

[22] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015. Translating Videos to Natural Language Using Deep Recurrent Neural Networks. In *NAACL-HLT*.

[23] L. Wang, Y. Li, J. Huang, and S. Lazebnik. 2019. Learning Two-Branch Neural Networks for Image-Text Matching Tasks. *TPAMI* (2019).

[24] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. 2019. Fine-Grained Action Retrieval Through Multiple Parts-of-Speech Embeddings. In *CVPR*. 450–459.

[25] Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, and Philipp Krahenbuhl. 2017. Sampling Matters in Deep Embedding Learning. In *ICCV*.

[26] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5987–5995.

[27] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*.

[28] Lin Xu, Han Sun, and Yuai Liu. 2019. Learning with batch-wise optimal transport loss for 3d shape recognition. In *CVPR*.

[29] Bowen Zhang, Hexiang Hu, and Fei Sha. 2018. Cross-Modal and Hierarchical Modeling of Video and Text. In *ECCV*.