

# Beyond Ranking Loss: Deep Holographic Networks for Multi-label Video Search

Anonymous

## Abstract

In this paper, we propose a Deep Holographic Networks (DHN) to learn similarity metrics of videos with multiple labels. DHN consists of 3 components: twin feature networks with shared architecture and weights, followed by a holographic compositional layer to interact the features of a video pair extracted from the feature networks, the output is fed to a stack of fully connected layers, with the top layer directly predicting similarity score of the video pair. The holographic composition layer has several favorable properties. First, instead of ranking loss driven deep metric learning (e.g. siamese loss and triplet loss), it explicitly encodes distance metric at intermediate layer of the network, which could potentially enable the network to learn richer pairwise relationships, e.g. the fine-grained similarity between multi-label videos. Moreover, it is parameter-free and memory efficient. We introduce a new large-scale video benchmark built upon the YouTube-8M dataset for the multi-label video search challenge task. Extensive evaluations on the benchmark dataset demonstrate that DHN performs significantly better than traditional deep metric learning approaches as well as other compositional networks.

## 1 Introduction

Content-based video search is to retrieve videos in a database that are the most similar to a query video. Feature representations and similarity metrics are the key components to a video search system. Previous work [Araujo and Girod, 2017] mainly concentrated on instance retrieval applications such as video copy detection and near-duplicate search, in which low-level features that can distinguish textured differences between fine-grained objects/scenes are developed, followed by standard distance computation between these features. These approaches did not consider high-level concepts depicted in videos, which are essential for semantic video search (e.g. event retrieval [Jiang *et al.*, 2007; Snoek *et al.*, 2009]) to close the so called semantic gap. This problem becomes more challenging when videos are associated with multiple labels, e.g. the number of truth labels per

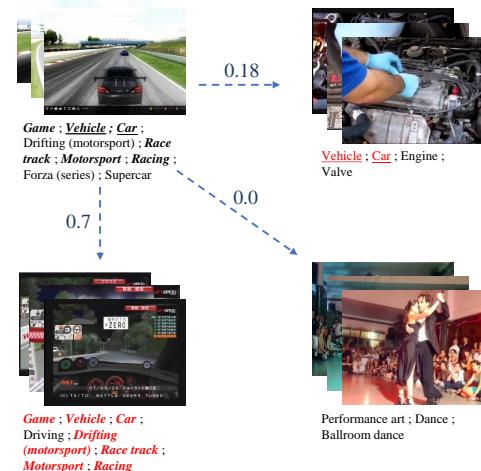


Figure 1: Examples of YouTube videos with coarse to fine-grained labels. Intuitively, the similarity metric between a video pair shall depend on the overlap ratio of their labels. The higher the overlap ratio, the more similar they are.

video varies from 1 to 31 for the recently introduced YouTube-8M dataset [Abu-El-Haija *et al.*, 2016] (Figure 1). In this work, we focus on the joint learning of feature representations and similarity metrics, tailored for multi-label video search.

With the rising of deep learning in recent years, there has been attempts to jointly learn feature representations and similarity metrics by deep neural networks for matching problems, e.g. face verification [Chopra *et al.*, 2005; Schroff *et al.*, 2015] and image instance retrieval [Radenović *et al.*, 2016]. The idea is to learn feature embedding of input samples to minimize certain loss function computed by distance of the embedded representations. Typically, feature embedding networks for videos are built upon Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM), which achieved remarkable success in applications like video classification [Karpathy *et al.*, 2014; Yue-Hei Ng *et al.*, 2015; Abu-El-Haija *et al.*, 2016] and activity recognition [Heilbron *et al.*, 2015]. However, metric learning with deep neural networks in video domain has not been well explored yet.

Currently, the mainstream in deep metric learning is de-

signing **ranking loss functions** on the top layer to optimize the similarity metrics. The first milestone work is Siamese Network [Chopra *et al.*, 2005] (see Figure 2(a)), a pairwise ranking loss termed contrastive loss is designed, with the objective to minimize the absolute distance between a matching pair and maximize the absolute distance of a non-matching pair. [Wang *et al.*, 2014] introduced Triplet Network by extending the network input from a pair to a triplet (i.e. a query, a positive and a negative sample), correspondingly, a triplet loss is defined to ensure that the distance between query and positive is smaller than the distance between query and negative with a pre-defined margin. Many other novel ranking loss functions have been proposed for further improvements of either siamese or triplet loss [Kumar *et al.*, 2016; Oh-Song *et al.*, 2016].

In this work, our first contribution is, instead of ranking loss driven deep metric learning, we propose an alternative solution by introducing the holographic composition [Nickel *et al.*, 2016]<sup>1</sup> as an additional layer after the feature embedding network (see Figure 2(b)). The idea of holographic composition is to enable direct interactions of features through either circular correlation or circular convolution [Plate, 1995]. The output of holographic composition layer is fed to a stack of fully connected layers, with the top layer directly predicting similarity score of a input pair. Holographic composition has several favorable properties. First, compared to siamese or triplet loss, it explicitly encodes distance metric (e.g. circular correlation in Equation 2) at intermediate layer of the network, which could potentially enable the network to learn richer pairwise relationships (e.g. the fine-grained similarity between multi-label videos in Figure 1). Second, the holographic composition layer is parameter-free. Third, holographic composition is memory efficient in the sense that it does not change the dimensionality of input to the bottom fully connected layer, while other compositional operators (e.g. tensor product) dramatically increase the number.

Our second contribution is, for the first time, we introduce the multi-label video search challenge task. To this end, we introduce a new large-scale multi-label video benchmark built upon the YouTube-8M dataset. We conduct systematic empirical evaluations of the proposed Deep Holographic Networks (DHN) over traditional deep metric learning approaches as well as the other compositional networks. Our observations on the new dataset demonstrate that DHN outperforms state-of-the-art by a large margin.<sup>2</sup>

## 2 Related Work

Traditional video instance search systems [Araujo and Girod, 2017; Lin *et al.*, 2017] are mainly built upon handcrafted features including local (e.g. SIFT [Lowe, 2004]) and global descriptors (VLAD [Jégou *et al.*, 2010] and FV [Perronnin *et al.*, 2010]), followed by standard distance computations. Recently, MPEG started the standardization of Compact Descriptors for Visual Analysis (CDVA) [Lin *et al.*, 2017], with

<sup>1</sup>This operator is closely related to the holographic models of associative memory.

<sup>2</sup>Source codes and the new dataset for multi-label video search are available at <http://double.blind>

the aim to come up with a normative bitstream of standardized features for mobile visual search and augmented reality applications.

In recent years, deep learning has also been developed for video applications [Xu *et al.*, 2015; Lin *et al.*, 2017]. [Xu *et al.*, 2015] proposed FV and VLAD aggregation techniques over dense local features of CNN activation maps from video frames for event retrieval. [Lin *et al.*, 2017] generated video representations by performing pooling over CNN activation maps extracted from video keyframes, which are subsequently used for video instance retrieval of objects, scenes and landmarks. LSTM has also been applied over frame-level CNN activations [Karpathy *et al.*, 2014] for modeling spatial-temporal clues [Yue-Hei Ng *et al.*, 2015]. Recent work on large scale video classification challenges such as ActivityNet [Heilbron *et al.*, 2015] and YouTube8M [Abu-El-Haija *et al.*, 2016] shown that the combination of CNN and LSTM obtains the best accuracy than simple aggregation techniques over CNN activations. In this work, we use the combined architecture of CNN and LSTM as the backbone network for learning video representations.

A number of recent works demonstrated that feature representation together with similarity metric can be learned in a unified deep learning framework [Li and Tang, 2015; Oh-Song *et al.*, 2016; Li and Tang, 2017]. [Chopra *et al.*, 2005; Radenović *et al.*, 2016] proposed siamese network to learn similarity metrics with contrastive loss for face verification and image instance retrieval, respectively. There are also triplet-based metric learning methods [Wang *et al.*, 2014; Schroff *et al.*, 2015; Arandjelovic *et al.*, 2016] that extends siamese network with triplet input {query, positive, negative}. [Han *et al.*, 2015] proposed to learn patch-level image descriptors by integrating a feature network and a metric network together.

Finally, our approach draws inspirations from recent work on holographic composition for link prediction on knowledge graph [Nickel *et al.*, 2016]. Holographic composition has also demonstrated its effectiveness for question answering in natural language processing [Socher *et al.*, 2013; Tay *et al.*, 2017]. However, to our knowledge, this work is the first to exploit holographic composition for deep metric learning in computer vision.

## 3 Deep Holographic Networks

### 3.1 Overview

As shown in Figure 2(b), the proposed deep holographic networks (DHN) can be decomposed into three key components: First, twin feature networks with shared architecture and weights for video-level feature extraction of a video pair independently. Following [Abu-El-Haija *et al.*, 2016], we use Inception network to extract CNN features for frames sampled from a video. Subsequently, the frame-level features are input to a LSTM network, and aggregated into a video-level feature representation with temporal feature encoded. In addition, multiple LSTMs can be stacked together to explore longer term dependencies of video frames. Second, the pair of video-level features are transformed to a new compositional feature representation by holographic composition layer, which we

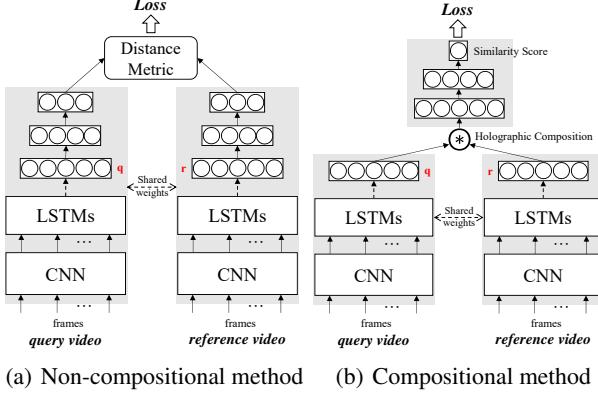


Figure 2: (a) Siamese Network (Non-compositional method) and (b) compositional network architectures for learning similarity metrics.

introduced in the following section. Finally, the compositional feature vector is fed to a stack of fully connected layers, and end with a sigmoid layer which directly predicts a similarity score of the video pair.

### 3.2 Holographic Composition

Instead of learning similarity metrics with ranking loss computed by distances of video pairs with representations output by the top layer, an alternative way for modeling similarity metrics is to explicitly interact activations of the pair at the intermediate holographic composition layer. Let  $\mathbf{q}$  and  $\mathbf{r}$  denote video-level feature vectors respectively extracted from query and reference videos using the twin feature networks. the ultimate goal is to predict the similarity score  $s(\mathbf{q}, \mathbf{r})$  of the pair with

$$s(\mathbf{q}, \mathbf{r}) = \sigma(\mathbf{W}^T(\mathbf{q} \circ \mathbf{r}) + b), \quad (1)$$

where  $\circ$  denotes a compositional function operated on the pair  $\{\mathbf{q}, \mathbf{r}\}$ , resulting in a compositional feature vector as input to the subsequent fully connected layers.  $\mathbf{W}$  and  $b$  denote learnable weights and bias of fully connected layers (for simplicity, Equation 1 contains only one fc layer),  $\sigma$  is the sigmoid function. Particularly, holographic composition can be implemented with either circular correlation or circular convolution [Plate, 1995].

**Circular correlation.** Let  $\circledast : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$  denotes the compositional operator of circular correlation, which is computed as

$$[\mathbf{q} \circledast \mathbf{r}]_k = \sum_{i=0}^{d-1} \mathbf{q}_i \mathbf{r}_{(k+i) \bmod d}, \quad k \in [0, d-1] \quad (2)$$

Figure 3(a) shows an example of circular correlation. Basically, circular correlation performs pairwise multiplications followed by summation with certain patterns. This operation can be interpreted as a compression of tensor product which enables rich representational learning without severely increasing the number of parameters of the network. In addition, the computation of Equation 2 can be significantly accelerated with fast Fourier transform (FFT) and inverse FFT, resulting in computational complexity of  $\mathcal{O}(d \log d)$ .

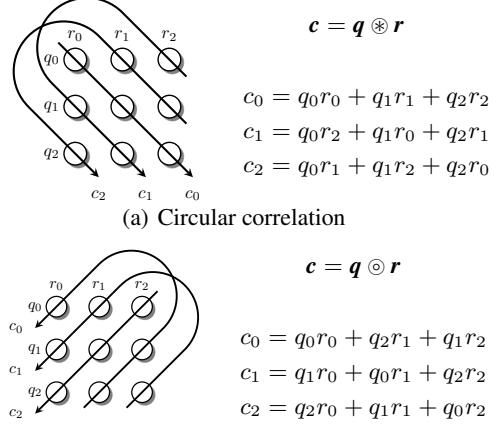


Figure 3: Holographic composition with either (a) circular correlation or (b) circular convolution. Circular arrow denotes summation operation.

**Circular convolution.** As shown in Figure 3(b), circular convolution is closely related to circular correlation,

$$[\mathbf{q} \odot \mathbf{r}]_k = \sum_{i=0}^{d-1} \mathbf{q}_i \mathbf{r}_{(k-i) \bmod d}, \quad k \in [0, d-1] \quad (3)$$

where  $\odot$  denotes the compositional operator of circular convolution.

The main differences between circular correlation and circular convolution are two-fold. First, circular correlation is non-commutative, i.e.,  $\mathbf{q} \circledast \mathbf{r} \neq \mathbf{r} \circledast \mathbf{q}$ , while circular convolution is commutative. Second, as shown in Figure 3(a), the first component of the compositional representation from circular correlation,  $[\mathbf{q} \circledast \mathbf{r}]_0$  (i.e.  $c_0$ ), represents the dot product of  $\mathbf{q}$  and  $\mathbf{r}$ , which closely relates to the cosine similarity of the pair, which is preferred for video search.

### 3.3 Loss function

Intuitively, the similarity of multi-label videos shall depend on the overlap ratio of their labels. In view of this, we compute the ground-truth semantic similarity between query and reference videos via Jaccard index,

$$\hat{s} = \jmath(q, p) = \frac{|q \cap p|}{|q \cup p|} \quad (4)$$

where  $q$  and  $p$  denotes the set of labels from query and reference videos. One may note that the holographic composition layer is parameter-free. Thus, we jointly train the twin feature networks and fully connected layers end to end by Stochastic Gradient Descent, with the objective to minimize either the simple mean squared error loss or the cross-entropy loss  $-(\hat{s} * \log(s(\mathbf{q}, \mathbf{r})) + (1 - \hat{s}) \log(1 - s(\mathbf{q}, \mathbf{r})))$ .

### 3.4 Inference

We apply the DHN for the task of multi-label video search. Given a query video, the goal is to rank the reference videos by their similarity scores to the query, which are directly output by the DHN for each video pair. With  $M$  queries and  $N$  reference

videos, the number of video pairs  $\{q, r\}$  is  $M \times N$ . Performing a forward pass through the DHN for each pair is ineffective as there are duplicated computations for feature networks. Instead, one can significantly accelerate the inference speed over multiple queries with a two-stage retrieval pipeline. First, video-level feature vectors are extracted with a single forward pass through the feature network for all the  $M + N$  videos. Second, the holographic composition layer and fully connected layers are used to generate similarity score for each of the  $M \times N$  pairs.

### 3.5 Discussions

#### DHN vs. Siamese network

Siamese network can be modeled by embedding a pair of samples into a low-dimensional Euclidean space independently, followed by standard distance comparison,

$$s(q, r) = \cos(f(q), f(r)) \quad (5)$$

where  $f(x) = \sigma(W^T x + b)$ ,  $\sigma$  is the sigmoid function,  $\cos(\cdot, \cdot)$  denotes cosine similarity. The parameters are optimized by minimizing the contrastive loss  $-(y * s(q, r) + (1 - y)\max(0, m - s(q, r)))$ , where  $y = 1$  when  $\{q, r\}$  is match, otherwise,  $y = 0$ .  $m$  is a constant margin.

First, the main difference of our DHN and siamese network is that the former explicitly interacts features of a pair at intermediate layer of the network architecture, while the latter does not. We argue that encoding metric at intermediate layer could potentially enable the network to learn richer pairwise relationships. Second, as shown in Table 1, the memory complexity of DHN is much smaller than siamese network, while with very small computational overhead  $\mathcal{O}(d \log d)$  introduced by circular correlation. Finally, DHN is storage free as it directly predicts similarity score, but siamese network needs to store feature representations output by the top layer. Due to limited space, we exclude triplet network from the discussion as its inference complexities are the same with siamese network.

#### DHN vs. Other compositional networks

**Deep Concatenation Network (DCN).** A straightforward approach for feature composition is to directly concatenate features from a pair of samples [Han *et al.*, 2015]. Let  $\oplus : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{2d}$  denotes concatenation operator, one may note that concatenation operator doubles the dimensionality the compositional vector, i.e. from  $d$  from  $2d$ . Thus, it increases the memory and computational cost of the fully connected layer by a factor of 2.

**Neural Tensor Network (NTN).** Concatenation does not interact features of a pair, another compositional operator termed tensor product is proposed [Socher *et al.*, 2013], which is in accordance with the outer product. An exhaustive pairwise multiplications between  $q$  and  $r$  is performed,

$$[q \otimes r]_{ij} = q_i r_j, \quad i, j \in [0, d - 1] \quad (6)$$

where  $\otimes$  denotes the tensor product operator. In this sense, tensor product allows to capture all pairwise interactions, with the cost of dramatically increased dimensionality of the compositional vector (from  $d$  to  $d^2$ ). This significantly increases the memory and computational cost of the subsequent fully connected layers. Also, It is worth noting that vanilla NTN

Table 1: Theoretical memory complexity of different methods. We only count the number of parameters for fully connected layers after the twin feature networks, by assuming there is only one hidden fc layer with  $h$  neurons and bias is not counted (see Figure 2). For Siamese/Triplet Network,  $t$  represents the number of neurons at the top layer ( $t >> 1$ ).

Network	# Parameters
Siamese / Triplet Network	$(d + t)h$
Deep Concatenation Network	$(2d + 1)h$
Neural Tensor Network	$(d^2 + 2d + 1)h$
DHN (Ours)	$(d + 1)h$

usually combines concatenation and tensor product together. Thus, the resulted compositional vector is with dimension of  $(d^2 + 2d)$ .

DHN differs from DCN in that holographic composition directly interacts features, instead of simple concatenation of them. On the other hand, holographic composition generates features with half the size of concatenation, resulting in less memory and computational complexity by a factor of 2 (see Table 1). A major advantage of DHN over NTN is that DHN enables modeling interactions of features, without significantly increasing the number of parameters of the subsequent layers (see Table 1).

## 4 Experiments

### 4.1 Dataset

The **YouTube-8M** video dataset [Abu-El-Haija *et al.*, 2016] contains around 8 million multi-label videos categorized into 4,716 classes. The number of ground truth labels per video varies from 1 to 31, with an average of 3.4 per video. This dataset presents two major challenges: diversity and class imbalance. The videos cover many different possible topics (music, politics, etc.), styles (CGI, documentary, etc.), and formats (conversation, action, etc.). Furthermore, there is significant class imbalance, with only 10 labels accounting for over half of all label appearances in the dataset, while a large portion of classes has a very small number of videos.

Class imbalance in queries would cause the evaluation accuracy driven by majority classes. To encourage a fair comparison, we build a new dataset termed **YouTube-MLR** designed for the multi-label video search task, based on the YouTube-8M set. We rank the classes by the number of videos in descending order, choose the top 1,000 classes, the rest classes are removed. As YouTube-8M ground-truth are only available for training and validation sets, we create the YouTube-MLR training and test set from the original YouTube-8M training and validation sets, respectively. In particular, the YouTube-MLR test set is constructed by sampling videos from the trimmed YouTube-8M validation set, randomly and evenly for each of the 1,000 classes selected. There is no duplicate video in YouTube-MLR test set.

To evaluate retrieval accuracy, the YouTube-MLR test set is further split into query and reference subsets. Furthermore, in order to evaluate performance trend at different scales, we generate a small-scale and a large-scale test set termed **YouTube-**

Table 2: Training and test dataset statistics.

	# Classes	Training	Validation	
YouTube-8M	4,716	4.906M	1,401,828	
	Test			
	Classes	Training	# Query	# Refs
Ours (small)	1,000	4.776M	4,569	20,133
Ours (large)			9,526	187,442

**MLR-S** and **YouTube-MLR-L**, respectively. Table 2 summarizes the statistics of YouTube-8M and the new YouTube-MLR-S/YouTube-MLR-L data sets.

## 4.2 Evaluate Metric

Normalized Discounted Cumulative Gain (NDCG) is a standard evaluation metric of ranking quality in information retrieval community, which takes the similarity level into consideration. NDCG is calculated as

$$NDCG@p = \frac{DCG@p}{IDCG@p} \quad (7)$$

where  $DCG@p = \sum_{i=1}^p \frac{2^{\hat{s}_i} - 1}{\log(i+1)}$  and  $IDCG@p = \sum_{i=1}^p \frac{2^{s_i} - 1}{\log(i+1)}$ .  $p$  is the truncated rank position;  $s_i$  and  $\hat{s}_i$  are the ground-truth similarity score and the predicted similarity score, respectively for the  $i$ -th position in a ranking list. For all experiments, we set  $p = \{1, 10, 100, 1000\}$  and compute mean NDCG (mNDCG) for all queries.

## 4.3 Model Training

It's impractical to process the hundreds of Terabytes YouTube-8M videos [Abu-El-Haija *et al.*, 2016]. Therefore, the organizers released pre-extracted features for video frames (1 frame per second, at most 300 frames per video), by extracting 2048-d feature vectors from the Inception network pre-trained on ImageNet, followed by PCA whitening to reduce feature dimension to 1024. In this work, we regard the 1024-d features as the CNN feature representations. The CNN features serve as the input to a two-layer LSTMs with 1024 hidden nodes. The output of LSTMs, i.e. concatenation of hidden states and cell states of each LSTM layer, is a 4096-d video-level feature vector as the input to the holographic composition layer.

To learn parameters of the layers after Inception network, we sample matching and non-matching video pairs from the YouTube-MLR training dataset. For each epoch, we randomly sample 10,000 from the training data as queries and 400 video pairs for each query, with the criterions that (1) the sampled queries are distributed evenly across classes; (2) query and positive videos are matching pair only if they have at least 1 label overlapped, otherwise, they are non-matching pair; (3) for each query, 60% of the 400 video pairs are matching pairs, the rest are non-matching pairs. Batch size is 100, 60% of them are matching pairs. We use learning rate 0.001, which is divided by 5 for every 10 epochs. To accelerate the convergency, we initialized the parameters of LSTMs with the model pre-trained on YouTube-8M [Wang *et al.*, 2017]. We train the network for 100 epochs with the Adam optimizer.

The inference speed for a video pair is around 55 ms on a single NVIDIA Tesla K40m.

In the following section, we refer DHN with circular correlation and circular convolution to **DHN-COR** and **DHN-CON**, respectively. We compare DHN to other compositional and non-compositional approaches, including Deep Concatenation Network (**DCN**) [Han *et al.*, 2015], Neural Tensor Network (**NTN**) [Socher *et al.*, 2013], **Siamese Net** [Radenović *et al.*, 2016] and **Triplet Net** [Arandjelovic *et al.*, 2016]. We tune hyper-parameters (e.g. architectures for the fully connected layers, margin for Siamese/Triplet Net, etc) to explore the performance trade-offs for each network architecture. At last, we also include a **Baseline** method by applying sum pooling over the 1024-d CNN features to form a new 1024-d video-level features. The similarity between a video pair is computed as the cosine similarity of their 1024-d video features.

## 4.4 Results

**DHN vs. Other compositional networks.** Table 3 shows comprehensive comparisons of DHN over other compositional variants, in terms of mNDCG on the YouTube-MLR-S test set. All compositional networks are trained with mean square error loss. We only count projection matrix of the FC layers, bias is ignored as it is insignificant. First, DHN obtains the best mNDCG@N scores across different rank positions (N=1, 10, 100, 1000), with significantly lower memory complexity over the rest. Compositional networks are superior to the baseline as N increases. Second, as expected, DHN-COR outperforms DHN-CON, suggesting that the circular correlation is more favored than circular convolution for the matching problem. Third, we study the effect of FC layer structures for compositional operators. There are consistent performance improvements if marginally increasing the number of FC parameters, the retrieval accuracy of DCN and NTN tends to drop if FC layers are too large, which is probably due to overfitting. Finally, mNDCG scores for the baseline are stable as N changes. However, in most cases, mNDCG scores increases with N for all compositional networks.

**Effect of Loss functions.** Table 4 studies the effect of loss functions (mean square error and cross entropy loss) for compositional networks, in terms of mNDCG@1000 on the YouTube-MLR-S test set. For simplicity, the FC layer structure for DCN, NTN, DHN-CON and DHN-COR follows the best performing model in Table 3, respectively. We observe that DCN with mean square error performs much worse than DCN with cross entropy, while NTN prefers mean square error. DHN-CON and DHN-COR are robust to both loss functions.

**DHN vs. Siamese/Triplet network.** Table 5 compares DHN to non-compositional methods including siamese network and triplet network on the YouTube-MLR-S test set. One can see that the DHN variants significantly outperform siamese and triplet networks, while with smaller memory complexity of FC layers. It's worth noting that the best performing triplet network is much better than the best performing siamese network (i.e. 0.638 vs. 0.562).

**How does DHN help?** One hypothesis is that holographic composition shall be helpful to boost the rank of {query, reference} pairs with higher overlap ratio of labels. To verify this, we compute the average rank position of reference videos

Table 3: Comparisons of the proposed DHN with other compositional methods on the YouTube-MLR-S test set.

Method	FC Layers	# Params	mNDCG@1	mNDCG@10	mNDCG@100	mNDCG@1000
Baseline	–	–	0.463	0.454	0.462	0.448
DCN	8K - 1	8.19K	0.059	0.050	0.078	0.133
	8K - 1K - 1	8.39M	0.341	0.387	0.496	0.616
	8K - 4K - 1K - 1	37.75M	0.255	0.298	0.404	0.589
NTN	16,785,408 - 2 - 1	33.57M	0.288	0.352	0.463	0.551
	16,785,408 - 4 - 1	67.14M	0.318	0.390	0.509	0.605
	16,785,408 - 6 - 1	100.71M	0.374	0.430	0.525	0.620
	16,785,408 - 8 - 1	134.28M	0.409	0.456	0.521	0.593
DHN-CON	4K - 1K - 1	4.20M	0.459	0.506	0.604	0.683
DHN-COR	4K - 1	4.10K	0.452	0.486	0.562	0.645
	4K - 1K - 1	4.20M	<b>0.527</b>	<b>0.563</b>	<b>0.628</b>	<b>0.691</b>

Table 4: Effect of loss functions, i.e. mean square error (MSE) and cross entropy loss, on compositional methods in terms of mNDCG@1000 with the YouTube-MLR-S test set.

Method	# Params	MSE	Cross Entropy
DCN	8.39M	0.616	0.670
NTN	100.71M	0.620	0.591
DHN-CON	4.20M	0.683	0.684
DHN-COR	4.20M	0.691	0.694

Table 5: Comparisons of the proposed DHN with non-compositional methods, on the YouTube-MLR-S test set.

Method	FC Layers	Margin	mNDCG@1000
Siamese	4K - 1K - 512	0.3	0.520
		0.5	0.535
		0.7	0.562
Triplet	4K - 1K - 512	0.3	0.585
		0.5	0.638
		0.7	0.596
DHN-CON	4K - 1K - 1	–	0.683
DHN-COR	4K - 1K - 1	–	0.691

as a function of their ground-truth similarity with queries (Equation 4), given the predict ranking list of the Baseline, Siamese Network and DHN-COR on the YouTube-MLR-S test set, respectively. With a collection of query-reference video pairs with ground-truth similarity ranging from 0.1 to 0.9 with step size 0.1, average rank position is computed as the median of the positions where reference videos ranked in the predict list. As shown in Figure 4(a), we observe that DHN-COR is able to bridge the gap between the Baseline/Siamese Network and “Ideal”, especially for lower ground-truth similarities. This means that DHN-COR is capable of ranking relevant reference videos with low overlap ratio of labels at top positions.

**Accuracy vs. Scale.** Finally, we study the performance trends of compositional networks as test data scales up. Figure 4(b) compares the performance loss when data scale increases from YouTube-MLR-S to YouTube-MLR-L, for DCN,

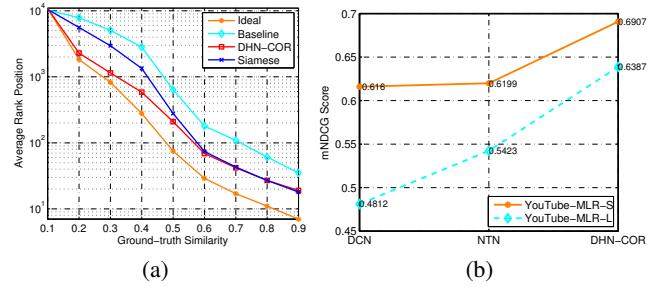


Figure 4: (a) Statistic of rank positions of reference videos as a function of their ground-truth similarity with queries, for the Baseline, Siamese Network and DHN-COR respectively. The higher the overlap ratio of labels between query and reference, the higher the ground-truth similarity. (b) Comparisons of retrieval performance trends between DHN and other compositional networks, as test dataset scales up.

NTN and DHN-COR respectively. As one can see, the relative performance loss of DHN-COR is the lowest (-7.54%, vs. NTN -12.53% and DCN -21.89%), implying that holographic composition is more robust to scale change than tensor product and concatenation operators.

## 5 Conclusion

In this paper, we propose deep holographic networks to learn similarity metrics of videos with multiple labels. We introduce holographic composition with circular correlation to explicitly model the interaction of a video pair with intermediate representations of the network architecture. This provides an alternative solution for deep metric learning, instead of the widely used ranking loss driven approaches built on the top layer (contrastive loss or triplet loss). In addition, the holographic operator is parameter-free and enables less memory footprint than state-of-the-art. We introduced a new video dataset built upon YouTube-8M for the task of multi-label video search. Promising results have been reported. Future work includes exploring DHN for other challenging applications such as image-caption and video-caption retrieval with images/videos annotated with rich text descriptions.

## References

- [Abu-El-Haija *et al.*, 2016] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.
- [Arandjelovic *et al.*, 2016] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, pages 5297–5307, 2016.
- [Araujo and Girod, 2017] Andre Araujo and Bernd Girod. Large-scale video retrieval using image queries. *IEEE Transactions on CSVT*, 2017.
- [Chopra *et al.*, 2005] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*, volume 1, pages 539–546. IEEE, 2005.
- [Han *et al.*, 2015] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *CVPR*, pages 3279–3286, 2015.
- [Heilbron *et al.*, 2015] FC Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.
- [Jégou *et al.*, 2010] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*, pages 3304–3311, 2010.
- [Jiang *et al.*, 2007] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *ACM international conference on Image and video retrieval*, pages 494–501, 2007.
- [Karpathy *et al.*, 2014] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, pages 1725–1732, 2014.
- [Kumar *et al.*, 2016] Vijay Kumar, Gustavo Carneiro, and Ian Reid. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *CVPR*, 2016.
- [Li and Tang, 2015] Zechao Li and Jinhui Tang. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE Transactions on Multimedia*, 17(11):1989–1999, 2015.
- [Li and Tang, 2017] Zechao Li and Jinhui Tang. Weakly supervised deep matrix factorization for social image understanding. *IEEE Transactions on Image Processing*, 26(1):276–288, 2017.
- [Lin *et al.*, 2017] Jie Lin, Ling-Yu Duan, Shiqi Wang, Yan Bai, Yihang Lou, Vijay Chandrasekhar, Tiejun Huang, Alex Kot, and Wen Gao. Hnip: Compact deep invariant representations for video matching, localization, and retrieval. *IEEE Transactions on Multimedia*, 19(9):1968–1983, 2017.
- [Lowe, 2004] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [Nickel *et al.*, 2016] Maximilian Nickel, Lorenzo Rosasco, Tomaso A Poggio, et al. Holographic embeddings of knowledge graphs. In *AAAI*, pages 1955–1961, 2016.
- [Oh-Song *et al.*, 2016] Hyun Oh-Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *CVPR*, pages 4004–4012, 2016.
- [Perronnin *et al.*, 2010] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *CVPR*, 2010.
- [Plate, 1995] Tony A Plate. Holographic reduced representations. *IEEE Transactions on Neural networks*, 6(3):623–641, 1995.
- [Radenović *et al.*, 2016] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European Conference on Computer Vision*, pages 3–20. Springer, 2016.
- [Schroff *et al.*, 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015.
- [Snoek *et al.*, 2009] Cees Snoek, Kvd Sande, OD Rooij, Bouke Huurnink, J Uijlings, M van Liempt, M Bugalhoj, I Trancosoy, F Yan, M Tahir, et al. The mediavill trecvid 2009 semantic video search engine. In *TRECVID workshop*, 2009.
- [Socher *et al.*, 2013] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in neural information processing systems*, pages 926–934, 2013.
- [Tay *et al.*, 2017] Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. Learning to rank question answer pairs with holographic dual lstm architecture. *arXiv preprint arXiv:1707.06372*, 2017.
- [Wang *et al.*, 2014] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, pages 1386–1393, 2014.
- [Wang *et al.*, 2017] Zhe Wang, Kingsley Kuan, Mathieu Ravaut, Gaurav Manek, Sibo Song, and et al. Truly multi-modal youtube-8m video classification with video, audio, and text. *arXiv preprint arXiv:1706.05461*, 2017.
- [Xu *et al.*, 2015] Zhongwen Xu, Yi Yang, and Alex G Hauptmann. A discriminative cnn video representation for event detection. In *CVPR*, pages 1798–1807, 2015.
- [Yue-Hei Ng *et al.*, 2015] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, pages 4694–4702, 2015.