BEYOND RANKING LOSS: DEEP HOLOGRAPHIC NETWORKS FOR MULTI-LABEL VIDEO SEARCH

Zhuo Chen[⋆] *Jie Lin*[†] *Zhe Wang*[†] *Vijay Chandrasekhar*[†] *Weisi Lin*[⋆]

[⋆] Nanyang Technological University, Singapore [†] Institute for Infocomm Research, A*STAR, Singapore

ABSTRACT

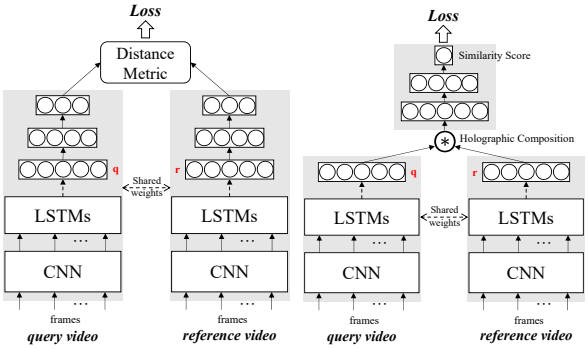
In this paper, we propose Deep Holographic Networks (DHN) to learn similarity metrics of videos for multi-label video search. DHN introduces a holographic composition layer to explicitly encode similarity metrics at intermediate layer of the network, instead of conventional deep metric learning approaches driven by ranking losses. The holographic composition layer is parameter-free and enables less memory footprint compared with state-of-the-art. Towards multi-label video search at large scale, we present a new video benchmark built upon the YouTube-8M dataset. Extensive evaluations on this dataset demonstrate that DHN performs better than traditional deep metric learning approaches as well as other compositional networks.

Index Terms— Video Search, Multi-label, Deep Metric Learning, Feature Composition

1. INTRODUCTION

Content-based video search is to retrieve videos in a database that are the most similar to a query video, in which low-level features [1, 2, 3] that can distinguish texture differences between fine-grained objects/scenes are developed, followed by standard distance computation between these features [4, 5]. These approaches did not consider high-level concepts depicted in videos, which are essential for semantic video search (e.g. event retrieval [6, 7]) to close the semantic gap. This problem becomes more challenging when videos are associated with multiple labels.

In visual retrieval systems, feature representations and similarity metrics are two key components. With the rising of deep learning in recent years, a number of recent works attempt to jointly learn feature representations and similarity metrics by deep neural networks, e.g. face verification [8, 9] and image instance retrieval [10]. However, metric learning with deep neural networks in video domain has not been well explored yet. Currently, the mainstream in deep metric learning is designing **ranking loss functions** on the top layer to optimize the similarity metrics. The first milestone work is Siamese Network [8] (see Figure 1(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. [11] 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 plus a pre-defined margin. Many other ranking loss functions have been proposed for further improvements of either siamese or triplet loss [12, 13].



(a) Non-compositional method (b) Compositional method

Fig. 1. (a) Siamese Network (Non-compositional method) and (b) compositional network architectures for metric learning.

Instead of deep metric learning approaches driven by ranking losses (Figure 1(a)) on top layer, we introduce holographic composition at intermediate layers to directly model pairwise video similarities (Figure 1(b)), inspired by recent work [14, 15] on holographic composition (termed by its close relation to holographic models of associative memory) for link prediction on knowledge graph and NLP. The idea of holographic composition is to enable direct interactions of features through either circular correlation or circular convolution [16]. The output of holographic composition layer is fed to a stack of fully connected layers, followed by Sigmoid layer that directly predict similarity score of a video pair. Holographic composition has several favorable properties. First, compared to siamese or triplet loss, it explicitly encodes similarity metrics (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).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.

Zhuo Chen and Jie Lin contribute equally to this paper.

To evaluate Deep Holographic Networks (DHN) in the context of large-scale multi-label video search, we introduce a new video benchmark built upon the YouTube-8M dataset [17]. We conduct systematic empirical evaluations of the proposed method 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.

2. DEEP HOLOGRAPHIC NETWORKS

2.1. Overview

As shown in Figure 1(b), the proposed deep holographic networks (DHN) can be decomposed into three key components: First, given a video pair, video-level features are extracted independently from twin feature networks with shared architecture and weights. Following [17], 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. Second, the pair of video-level features are transformed to a new compositional feature representation by holographic composition layer. 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 for the video pair.

2.2. Holographic Composition

Instead of learning similarity metrics with pre-defined ranking losses, an alternative way for modeling similarity metrics is to explicitly interact pair-wise video feature representations at intermediate layer via compositional operation. Assuming $\boldsymbol{q}$ and $\boldsymbol{r}$ are video-level features extracted from query and reference videos, the similarity score of the video pair can be generally inferred with

$$s(\boldsymbol{q}, \boldsymbol{r}) = \sigma(W^T(\boldsymbol{q} \bigcirc \boldsymbol{r}) + b), \tag{1}$$

where $\bigcirc$ denotes a compositional function operated on the pair $\{\boldsymbol{q}, \boldsymbol{r}\}$, resulting in a compositional feature vector as input to the subsequent fully connected layers. $\boldsymbol{W}$ and b denote learnable weights and bias of fully connected layers (for simplicity, Equation 1 contains only one fc layer), σ is the activation function. Particularly, holographic composition can be implemented with either circular correlation or circular convolution [16].

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

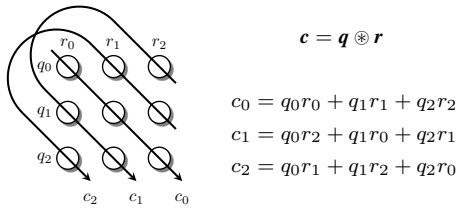
$$[\boldsymbol{q} \circledast \boldsymbol{r}]_k = \sum_{i=0}^{d-1} \boldsymbol{q}_i \boldsymbol{r}_{(k+i) \bmod d}, \quad k \in [0, d-1] \tag{2}$$

Figure 2(a) shows an example of circular correlation. Basically, circular correlation performs pairwise multiplications followed by summation with certain patterns. In addition, the computation of Equation 2 can be significantly accelerated with fast Fourier transform (FFT) and inverse FFT, resulting in computational complexity of $\mathcal{O}(d \log d)$.

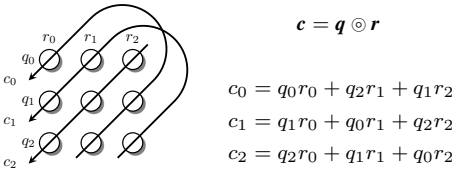
Circular convolution. As shown in Figure 2(b), circular convolution is closely related to circular correlation,

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

where $\circledcirc$ 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., $\boldsymbol{q} \circledast \boldsymbol{r} \neq \boldsymbol{r} \circledast \boldsymbol{q}$, while circular convolution is commutative. Second, as shown in Figure 2(a), the first component of the compositional representation from circular correlation, $[\boldsymbol{q} \circledast \boldsymbol{r}]_0$ (i.e. c_0), represents the dot product of $\boldsymbol{q}$ and $\boldsymbol{r}$, which closely relates to the cosine similarity of the pair, which is preferred for video search.



(a) Circular correlation



(b) Circular convolution

Fig. 2. Holographic composition with either (a) circular correlation or (b) circular convolution. Circular arrow denotes summation operation.

2.3. Discussions

2.3.1. 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,

$$s(\boldsymbol{q}, \boldsymbol{r}) = cos(f(\boldsymbol{q}), f(\boldsymbol{r})) \tag{4}$$

where $f(\boldsymbol{x}) = \sigma(W^T\boldsymbol{x} + b)$, σ is the sigmoid function, $cos(\cdot, \cdot)$ denotes cosine similarity. The parameters are optimized by minimizing the contrastive loss $-(y * s(\boldsymbol{q}, \boldsymbol{r}) + (1-y)max(0, m - s(\boldsymbol{q}, \boldsymbol{r})))$, where $y = 1$ when $\{\boldsymbol{q}, \boldsymbol{r}\}$ is matched, otherwise, $y = 0$. m is a constant margin. The main difference of DHN and siamese network is that DHN explicitly interacts paired features at intermediate layer of the network architecture. It embeds similarity metrics at intermediate layer, which could enable the network to learn richer pairwise relationships. Moreover, as shown in Table 1, the memory complexity of DHN is much smaller than siamese network.

2.3.2. 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 [18]. Let $\oplus : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{2d}$

Table 1. Theoretical memory complexity of different methods. Only the layers after video embeddings are counted, assuming there is only one hidden fc layer with h neurons and 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)	$(d+1)h$

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 (see Table 1).

Neural Tensor Network (NTN). Tensor product [19] performs exhaustive pairwise multiplications between vectors, which is in accordance with the outer product:

$$[\boldsymbol{q} \otimes \boldsymbol{r}]_{ij} = \boldsymbol{q}_i \boldsymbol{r}_j, \quad i,j \in [0, d-1]. \tag{5}$$

Tensor product allows to capture all pairwise interactions with the cost of dramatically increased dimensionality of the compositional vector from d to d^2. Vanilla NTN usually combines concatenation and tensor product together, which further enlarge the compositional vector size to (d^2+2d). Thus, 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).

3. EXPERIMENTS

3.1. Dataset

The **YouTube-8M** video dataset [17] 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. As class imbalance in queries would cause the evaluation accuracy driven by majority classes, to encourage a fair comparison, we tailor YouTube-8M to a new multi-label video retrieval dataset, termed **YouTube-MLR**. Concretely, the videos of the most frequent 1,000 classes are chosen to construct the new dataset. 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. To evaluate retrieval accuracy, the YouTube-MLR test set is further splitted into query and reference subsets. Moreover, in order to evaluate performance trend at different scales, we generate a small-scale and a large-scale test set termed **YouTube-MLR-S** and **YouTube-MLR-L**, respectively. Table 2 summaries the statistics of YouTube-8M and the new YouTube-MLR datasets.

In addition, the ground-truth similarity of a video pair is given by the overlap ratio (i.e. Jaccard index) of their labels,

$$\hat{s} = \jmath(q,p) = \frac{|q \cap p|}{|q \cup p|} \tag{6}$$

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

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	**0.527**	**0.563**	**0.628**	**0.691**

where q and p denotes the labels of query and reference videos.

3.2. 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

$$NDCG@p = \frac{DCG@p}{IDCG@p} \tag{7}$$

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.

3.3. Model Training

As it's impractical to process the hundreds of Terabytes YouTube-8M videos, we train our models on top of the frame-level features extracted by [17]. The frame-level features serve as the input of a two-layer LSTMs with 1024 hidden nodes. The output of LSTMs, i.e. video-level features, are passed to the holographic composition layer.

To generate the training batches, 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) sampled queries are distributed evenly across each label; (2) for each query, 60% of the 400 video pairs are matching pairs (similarity scores are non-zero), the rest are non-matching pairs. Batch size is set as 100, and learning rate is 0.001 which is divided by 5 for every 10 epochs. The LSTMs is initialized with the pre-trained model [20] to accelerate the convergency. The network is trained for 100 epochs with the Adam optimizer to minimize either the simple mean squared error loss or the cross-entropy loss.Training takes around one week on a single NVIDIA Tesla K40m,

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

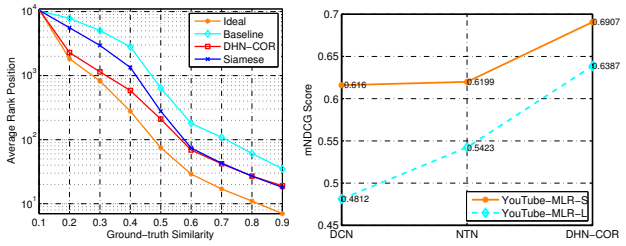
while inference speed for a video pair is around 55 ms.

We refer DHN with circular correlation and circular convolution to **DHN-COR** and **DHN-CON**, respectively. DHN is compared with other compositional and non-compositional approaches, including Deep Concatenation Network (**DCN**) [18], Neural Tensor Network (**NTN**) [19], **Siamese Net** [10] and **Triplet Net** [21]. We also include a **Baseline** method by applying sum pooling over the frame-level features to form the video-level feature, where the similarity between video pairs is induced by the cosine similarity of the video features.

3.4. Results

DHN vs. Other compositional networks. Table 3 shows comprehensive comparisons of DHN over other compositional variants. All compositional networks are trained with mean square error loss. From the results, DHN obtains the best mNDCG@N scores across different rank positions (N=1,10,100,1000), with significantly lower memory complexity over the rest. In particular, as expected, DHN-COR outperforms DHN-CON, suggesting that the circular correlation is more favored than circular convolution for the matching problem. Furthermore, 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.

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.



(a) (b)

Fig. 3. (a) Statistic of rank positions of reference videos as a function of their ground-truth similarity with queries. (b) Comparisons of retrieval performance trends between DHN and other compositional networks, as test dataset scales up.

DHN vs. Siamese/Triplet network. Table 5 compares DHN to non-compositional methods including siamese and triplet network on the YouTube-MLR-S test set. One can see that the DHN variants outperform siamese and triplet networks with smaller memory complexity for FC layers.

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 in retrieval database as a function of their ground-truth similarity with queries. As shown in Figure 3(a), we observe that DHN-COR is the closest to the ideal solution, 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 3(b) compares the performance loss when data scale increases from YouTube-MLR-S to YouTube-MLR-L.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.

4. CONCLUSION

In this paper, holographic composition was introduced in deep neural networks to explore the similarity metrics of video features at intermediate layers. This provides an alternative solution for deep metric learning, instead of the widely used ranking losses (contrastive or triplet loss) built on top layer. Promising results have been reported on a new large-scale multi-label video benchmark built upon the YouTube-8M dataset.

Acknowledgement This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University and the Agency for Science, Technology and Research (A*STAR) under its Hardware-Software Co-optimisation for Deep Learning (Project No.A1892b0026). This work is also partially supported by Singapore Ministry of Education Tier-2 Fund MOE2016-T2-2-057(S).

5. REFERENCES

[1] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[2] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010, pp. 3304–3311.

[3] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier, "Large-scale image retrieval with compressed fisher vectors," in *CVPR*, 2010.

[4] Andre Araujo and Bernd Girod, "Large-scale video retrieval using image queries," *IEEE Transactions on CSVT*, 2017.

[5] 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*, vol. 19, no. 9, pp. 1968–1983, 2017.

[6] 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*, 2007, pp. 494–501.

[7] Cees Snoek, Kvd Sande, OD Rooij, Bouke Huurnink, J Uijlings, M van Liempt, M Bugalhoy, I Trancosoy, F Yan, M Tahir, et al., "The mediamill trecvid 2009 semantic video search engine," in *TRECVID workshop*, 2009.

[8] Sumit Chopra, Raia Hadsell, and Yann LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *CVPR*. IEEE, 2005, vol. 1, pp. 539–546.

[9] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.

[10] 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*. Springer, 2016, pp. 3–20.

[11] 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*, 2014, pp. 1386–1393.

[12] 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.

[13] Hyun Oh-Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese, "Deep metric learning via lifted structured feature embedding," in *CVPR*, 2016, pp. 4004–4012.

[14] Maximilian Nickel, Lorenzo Rosasco, Tomaso A Poggio, et al., "Holographic embeddings of knowledge graphs.," in *AAAI*, 2016, pp. 1955–1961.

[15] 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.

[16] Tony A Plate, "Holographic reduced representations," *IEEE Transactions on Neural networks*, vol. 6, no. 3, pp. 623–641, 1995.

[17] 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.

[18] Xufeng Han, Thomas Leung, Yangqing Jia, Rahul Sukthankar, and Alexander C Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *CVPR*, 2015, pp. 3279–3286.

[19] 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*, 2013, pp. 926–934.

[20] 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.

[21] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *CVPR*, 2016, pp. 5297–5307.