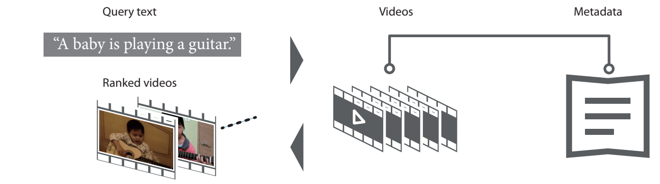
Learning Joint Representations of Videos and Sentences with Web Image Search

M. Otani[1], Y. Nakashima[1], E. Rahtu[2], J. Heikkilä[2], N. Yokoya[1]
[1] {otani.mayu.ob9, n-yuta, yokoya}@is.naist.jp Graduate School of Information Science, Nara Institute of Science and Technology
[2] {erahtu, jth}@ee.oulu.fi Center for Machine Vision and Signal Analysis, University of Oulu

Video retrieval by natural language queries



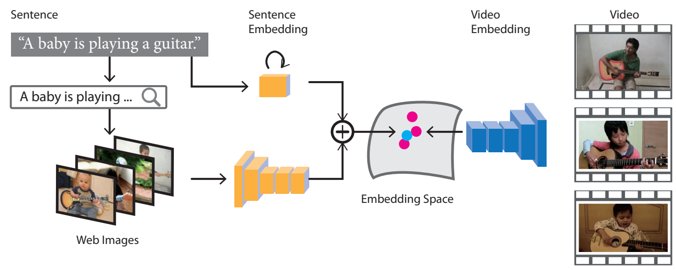
Difficulties in video retrieval by natural language queries

- User-generated metadata is unreliable
- Lack of metadata for videos

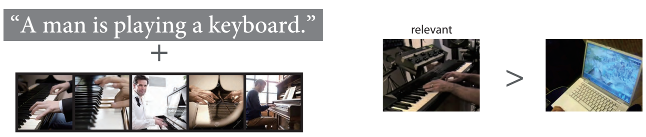
Goal: Retrieve videos relevant to a natural language query based on videos' content

Overview of our approach

Similarity estimation using joint embedding space

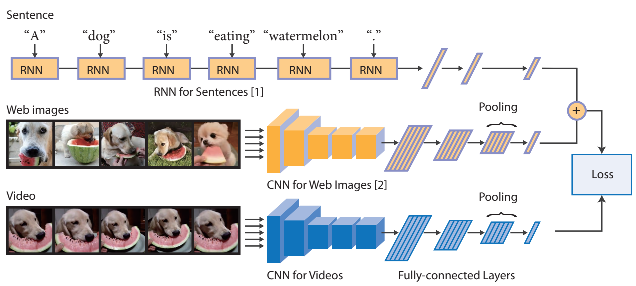


- Map semantically similar videos or sentences to close points
- Incorporate web images to disambiguate semantics in queries



Joint learning of embedding models

Model architechture



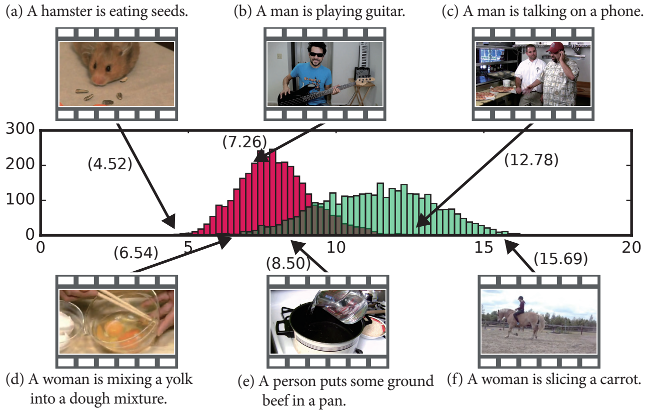
Contrastive loss

$$loss(X_n, Y_n) = t_n d(X_n, Y_n) + (1 - t_n) \max(\alpha - d(X_n, Y_n), 0)$$

$d(X_n, Y_n)$: Euclidean distance between video and sentence embeddings

t_n : Label. $t_n = 1$ if the video and the sentence is relevant,otherwise $t_n = 0$

Learned distance between videos and sentences



Video retrieval experiment

Dataset

- Microsoft Research Video Description Corpus [3]
- 1970 videos, 5 descriptions for each video

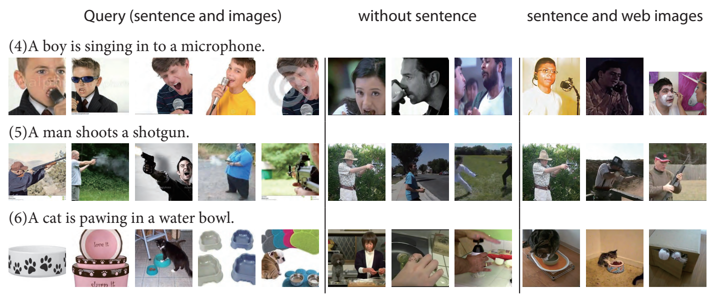
Video retrieval results

1. Comparison to a model trained without web images

Web images reduce the ambiguity of semantics in queries.

2. Comparison to a model trained without sentences

Both sentene and web images are necessary to compute embeddings



Video retrieval scores

- Video retrieval: find 1 correct video out of 670 videos
- Sentence retrieval: find one of 5 correct sentences out of 3350 videos

	Video retrieval					Sentence retrieval				
Models	R@1	R@5	R@10	aR	mR	R@1	R@5	R@10	aR	mR
Random Ranking	0.14	0.79	1.48	335.92	333	0.22	0.69	1.32	561.32	439
VGG+VS	6.12	21.88	33.22	58.98	24	7.01	18.66	27.16	131.33	35
VGG+VI	4.03	13.70	21.40	94.62	48	5.67	17.91	28.21	116.86	38
VGG+ALL$_1$	6.48	20.15	30.51	59.53	26	**10.60**	25.22	36.42	85.90	21
VGG+ALL$_2$	5.97	21.31	32.54	56.01	24	8.66	22.84	33.13	100.14	29
GoogLeNet+VS	7.49	22.84	33.10	54.14	22	8.51	21.34	30.45	114.66	33
GoogLeNet+VI	4.24	16.42	24.96	84.48	41	6.87	17.31	30.00	96.78	30
GoogLeNet+ALL$_1$	5.52	18.93	28.90	60.38	28	9.85	**27.01**	**38.36**	**75.23**	**19**
GoogLeNet+ALL$_2$	**7.67**	**23.40**	**34.99**	**49.08**	**21**	9.85	24.18	33.73	85.16	22
ST [1]	2.63	11.55	19.34	106.00	51	2.99	10.90	17.46	241.00	77
DVCT [4]	-	-	-	224.10	-	-	-	-	236.27	-

R@K: Recall at top-K results aR: Average rank mR: Median rank

Conclusion:

- We trained embedding of videos and sentences into a joint space for semantic similarity estimation.
- The use of web images disambiguates the visual concepts of query text.

Future work:

- Video embedding that considering temporal structures of videos

[1] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, "Skip-Thought Vectors," NIPS, pp. 3276–3284, 2015.
[2] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," CVPR, pp. 1–9, 2015.
[3] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," ACL' 11, pp. 190–200.
[4] R. Xu, C. Xiong, W. Chen, and J. Corso, "Jointly modeling deep video and compositional text to bridge vision and language in a unified framework," AAAI, pp. 2346–2352, 2015.