University of Amsterdam

Unified Embedding and Metric Learning for Zero-Exemplar Event Detection

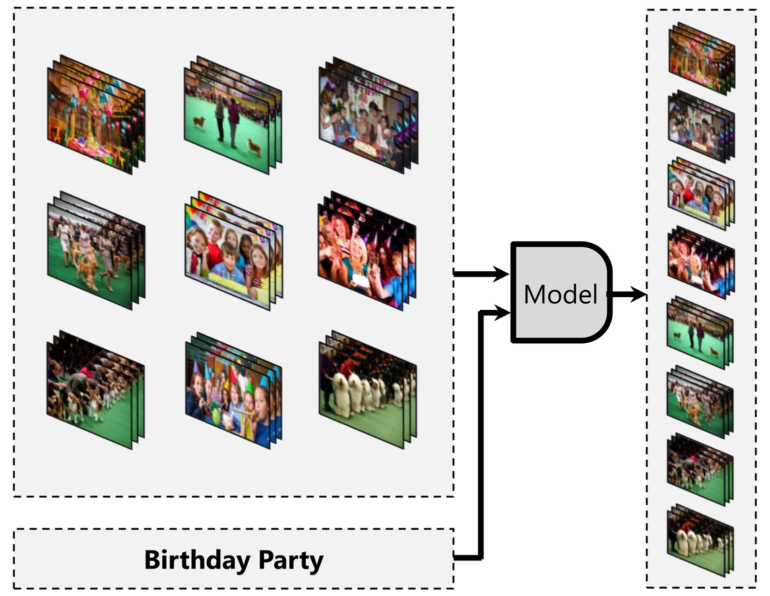
Noureldien Hussein, Efstratios Gavves, Arnold W. M. Smeulders | Quva Lab, University of Amsterdam

IEEE Conference on Computer Vision and Pattern Recognition

CVPR 2017

July 21-26, Honolulu

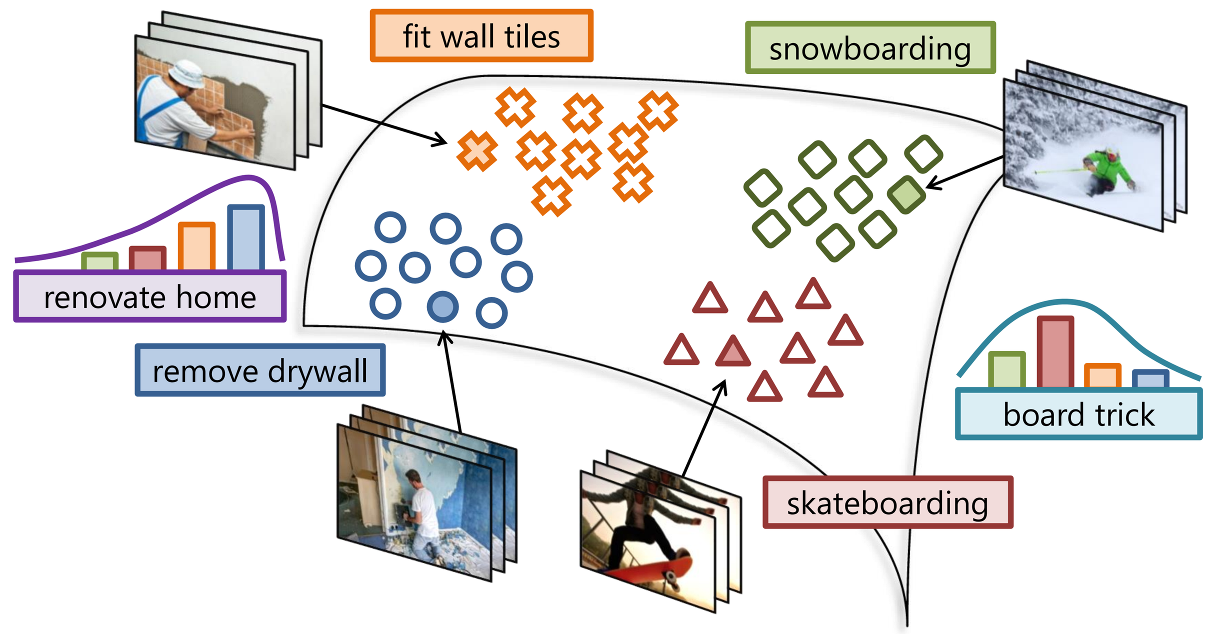
Problem



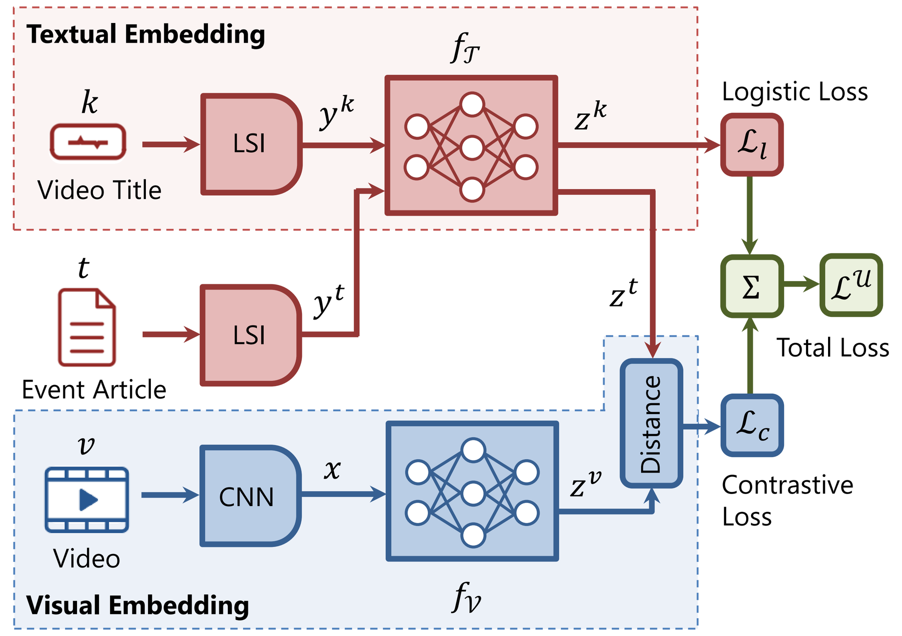
Zero-exemplar Event Detection (ZED) is posed as a video retrieval task. Given test videos and a novel query, the model is required to rank the videos accordingly.

We use video samples from EventNet and event articles from WikiHow.

Approach



We pose ZED as learning from a set of predefined events. Given video exemplars of events "removing drywall" or "fit wall tiles", one may detect a novel event "renovate home" as a probability distribution over the predefined events.



$$\mathcal{L}^{\mathcal{U}} = \mathcal{L}_{con} + \mathcal{L}_{log} \qquad (1)$$

$$\mathcal{L}_{log} = \sum_{i=1}^{N}\sum_{j=1}^{M} -y_i^j \log f_{\mathcal{T}}^j(k_i; W_{\mathcal{T}})$$

$$\mathcal{L}_{con} = \frac{1}{2N}\sum_{i=1}^{N} h_i \cdot d_i^2 + (1-h_i)\max(1-d_i, 0)^2$$

$$d_i = \|f_{\mathcal{T}}(t_i; W_{\mathcal{T}}) - f_{\mathcal{V}}(v_i; W_{\mathcal{V}})\|_2$$

At the top, network $f_{\mathcal{T}}$ learns to classify the title feature y^k into one of M event categories. In the middle, we borrow the network $f_{\mathcal{T}}$ to embed the event articles feature y^t as $z^t \in \mathcal{Z}$. Then, at the bottom, the network $f_{\mathcal{V}}$ learns to embed the video feature x as $z^v \in \mathcal{Z}$ such that the distance between (z^t, z^v) is minimized, in the learned metric space $\mathcal{Z}$.

Novelties

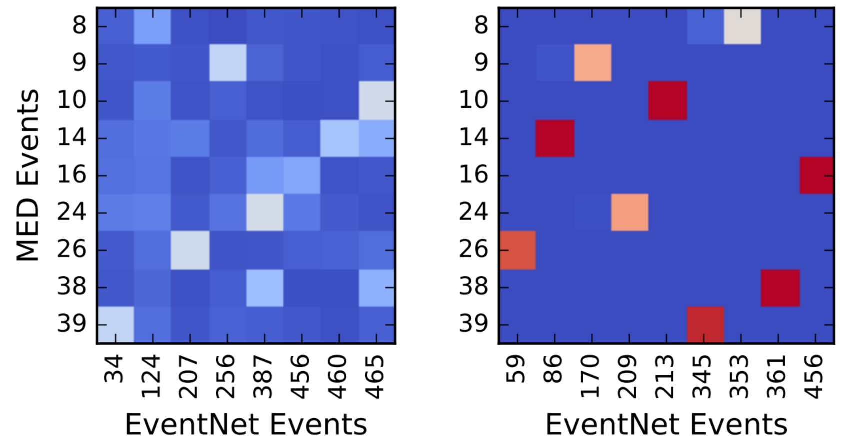
1 **Unified embedding** for cross-modalities with **metric loss** for maximum discrimination between events.

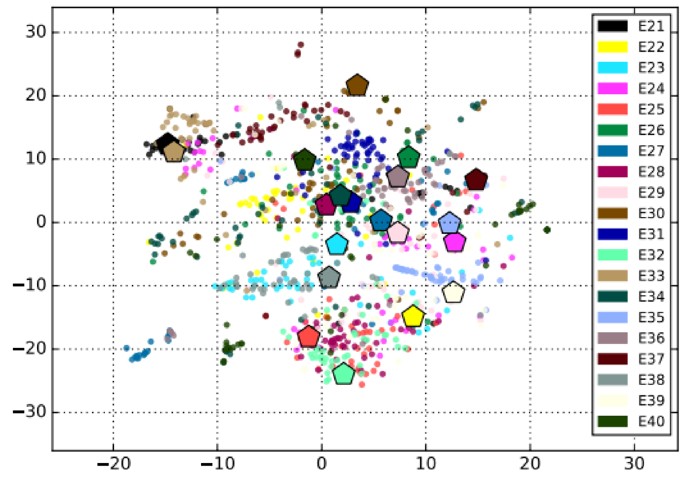
2 **Textual embedding** poses a novel query a probability of predefined events.

3 **External data source**, of event articles and related videos, with end-to-end learning from cross-modal pairs.

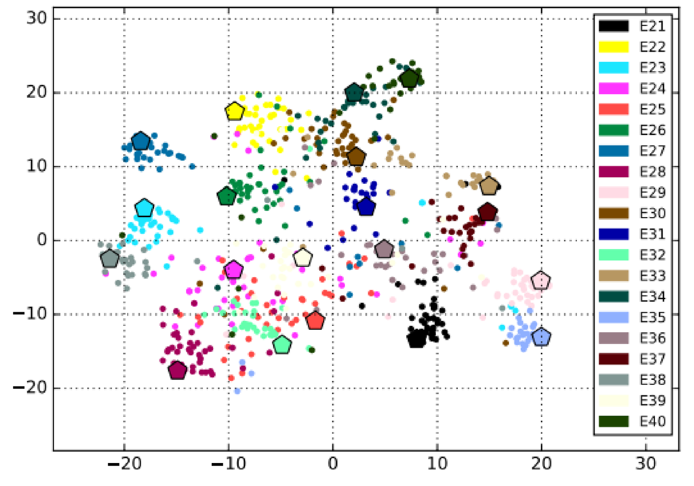
Results

Out textual embedding $f_{\mathcal{T}}$ maps the text description of MED events to EventNet events better than off-the-shelf LDA, LSI or Doc2vec.



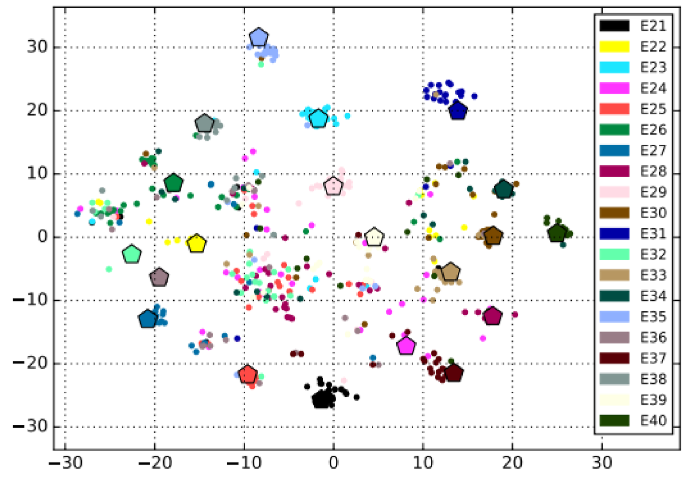


(a) Visual Embedding ($model^{\mathcal{V}}$)

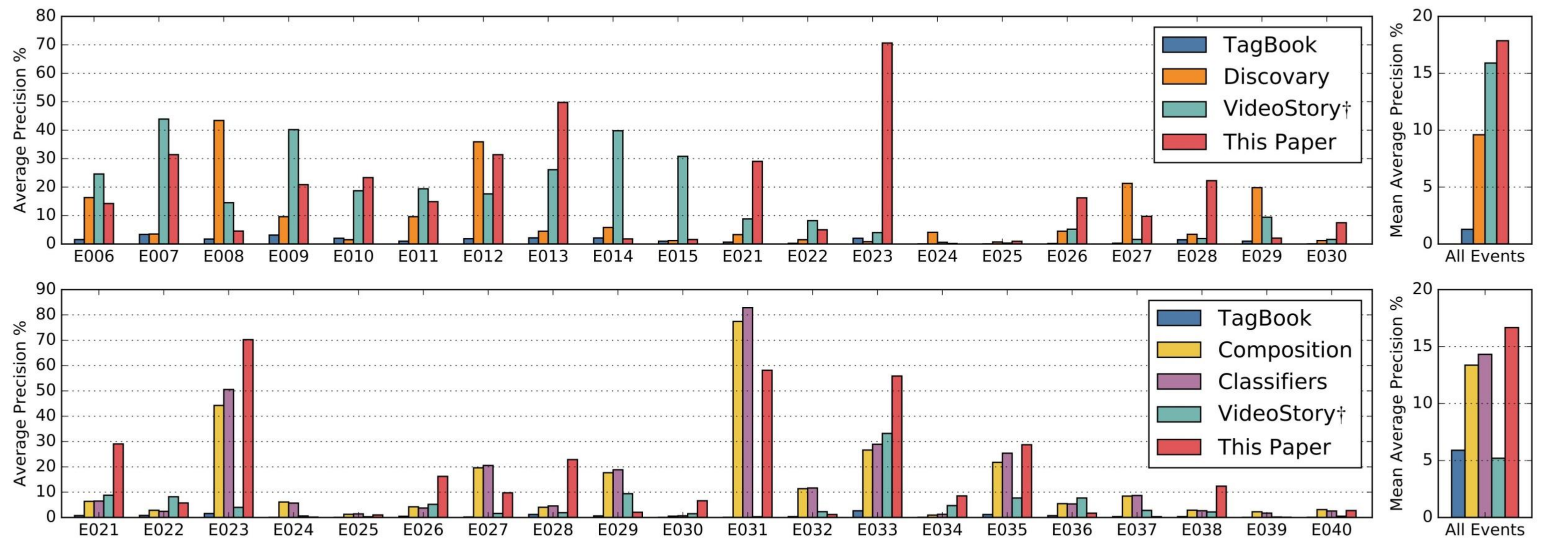


(b) Separate Embedding ($model^{\mathcal{S}}$)

- The unified embedding (c) is doing a better job in discriminating the text articles of the events.
- In (c), projecting the videos on their related events is much better than the baselines (a) and (b).



(c) Unified Embedding ($model^{\mathcal{U}}$)



Event detection accuracies: per-event average precision (AP) and per-dataset mean average precision (mAP) for MED-13 and MED-14 datasets.

Success (left) and failure (right) video examples of three different events: (a) birthday party, (b) dog show, (c) renovating home.

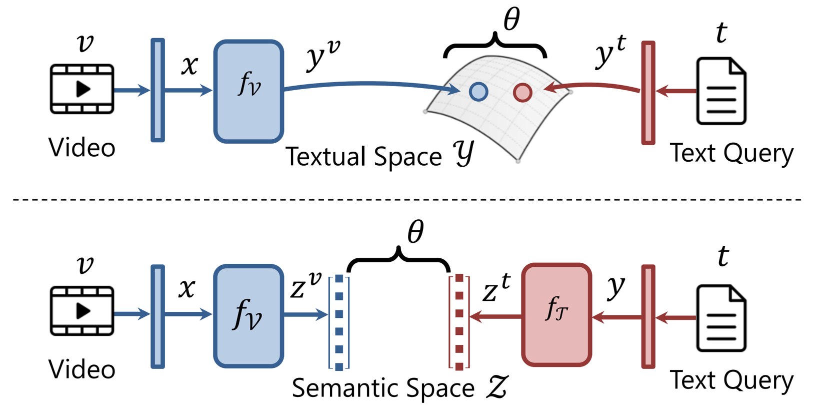
Method		MED13	MED14
TagBook [18]	ToM '15	12.90	05.90
Discovary [7]	ICAI '15	09.60	–
Composition [8]	AAAI '16	12.64	13.37
Classifiers [9]	CVPR '16	13.46	<u>14.32</u>
VideoStory† [17]	PAMI '16	15.90	05.20
VideoStory* [17]	PAMI '16	**20.00**	08.00
This Work ($model^{\mathcal{U}}$)		<u>17.86</u>	**16.67**

Baseline	Loss		Metric	$f_{\mathcal{V}}(\cdot)$	$f_{\mathcal{T}}(\cdot)$	MED13	MED14
$model^{\mathcal{V}}$	$\mathcal{L}_{mse}^{\mathcal{V}}$	(2)	✗	✓	✗	11.90	10.76
$model^{\mathcal{C}}$	$\mathcal{L}_{con}^{\mathcal{C}}$	(3)	✓	✓	✗	13.29	12.31
$model^{\mathcal{S}}$	$\mathcal{L}_{log}$	(4)	✗	✓	✓	15.60	13.49
$model^{\mathcal{N}}$	$\mathcal{L}_{mse}^{\mathcal{N}}$	(5)	✗	✓	✓	15.92	14.36
$model^{\mathcal{U}}$	$\mathcal{L}^{\mathcal{U}}$	(1)	✓	✓	✓	17.86	16.67

Left: retrieval accuracy (mAP) of our model vs. related works for MED-13 and MED-14 datasets. Right: retrieval accuracy (mAP) of our model (unified embedding) vs. other baselines.

Experiments

Model overview of baseline methods. Top: visual embedding ($model^{\mathcal{V}}$). Bottom: separate embedding ($model^{\mathcal{S}}$)



$$\mathcal{L}_{mse}^{\mathcal{V}} = \frac{1}{N}\sum_{i=1}^{N}\|y_i - f_{\mathcal{V}}(v_i; W_{\mathcal{V}})\|_2^2 \qquad (2)$$

$$\mathcal{L}_{con}^{\mathcal{C}} = \frac{1}{2N}\sum_{i=1}^{N} h_i \cdot d_i^2 + (1-h_i)\max(1-d_i, 0)^2 \qquad (3)$$

$$\mathcal{L}_{log} = \sum_{i=1}^{N}\sum_{j=1}^{M} -y_i^j \log f_{\mathcal{T}}^j(k_i; W_{\mathcal{T}}) \qquad (4)$$

$$\mathcal{L}_{mse}^{\mathcal{N}} = \frac{1}{N}\sum_{i=1}^{N}\|f_{\mathcal{T}}(t_i; W_{\mathcal{T}}) - f_{\mathcal{V}}(v_i; W_{\mathcal{V}})\|_2^2 \qquad (5)$$

Loss functions used to train the baseline models: visual embedding $model^{\mathcal{V}}$(2), contrastive visual $model^{\mathcal{C}}$(3), separate embedding $model^{\mathcal{S}}$(4) and non-metric embedding $model^{\mathcal{N}}$(5).

Take Home

The Good: external knowledge (EventNet, WikiHow) is leveraged for better zero-exemplar event detection.

The Bad: no fine-grained event detection, e.g. "fixing musical instrument" vs. "tuning musical instrument".

The Ugly: is average pooling enough for video representation or temporal modeling is required?

GitHub: https://git.io/vS2Cf