[ICCV 2019] ViSiL: Fine-grained Spatio-Temporal Video Similarity Learning

발표자 : 이현주

Instagram

Youtube

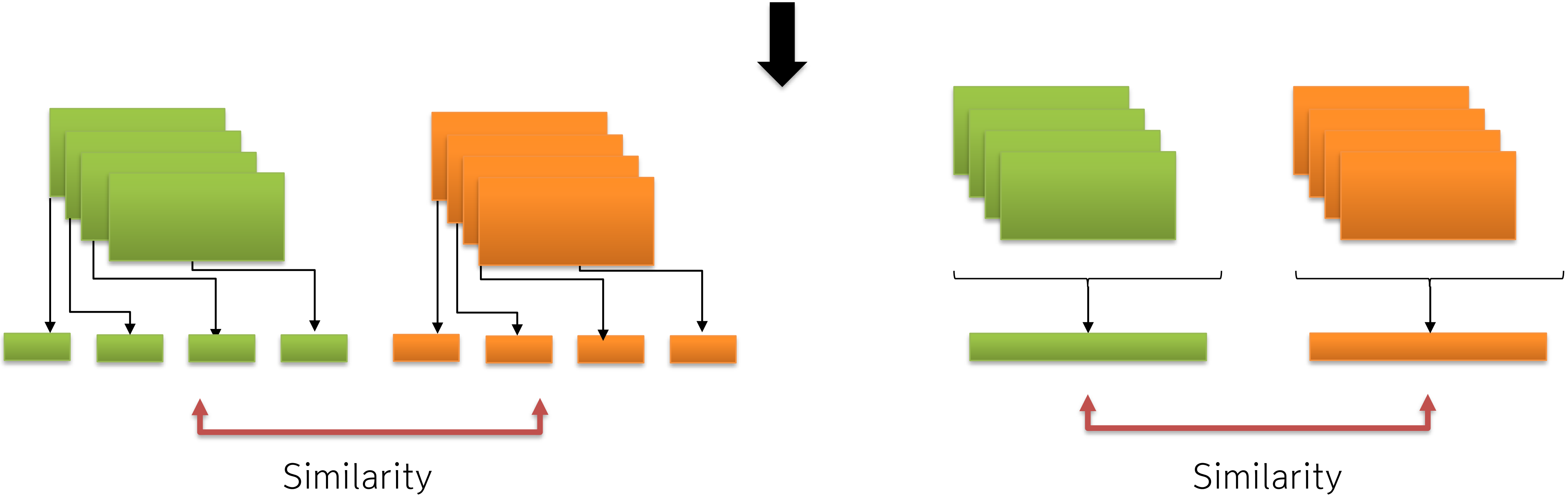
TikTok

the problem of similarity estimation between pairs of videos

Overview

previous video retrieval approaches
: embed the **whole frame** or the **whole video** into a vector descriptor before the similarity estimation

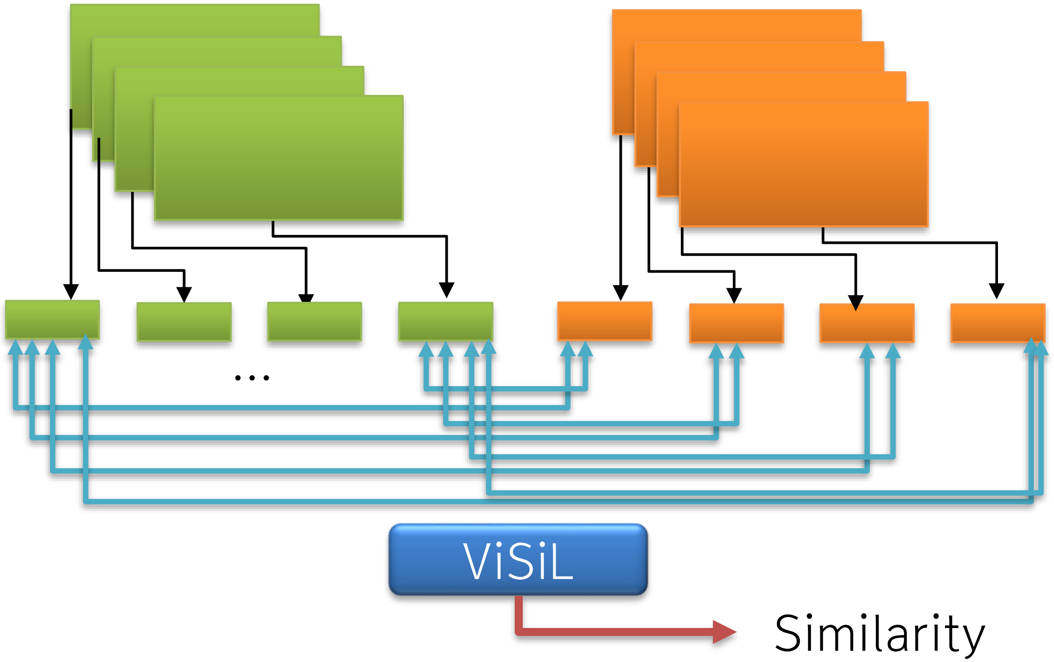
→ lost fine-grained Spatio-Temporal relations between pairs of videos



Overview

Our ViSiL approach
: train CNN-based approach to calculate **video-to-video similarity** from refined **frame-to-frame similarity** matrices

→ **consider both intra- and inter-frame relations**



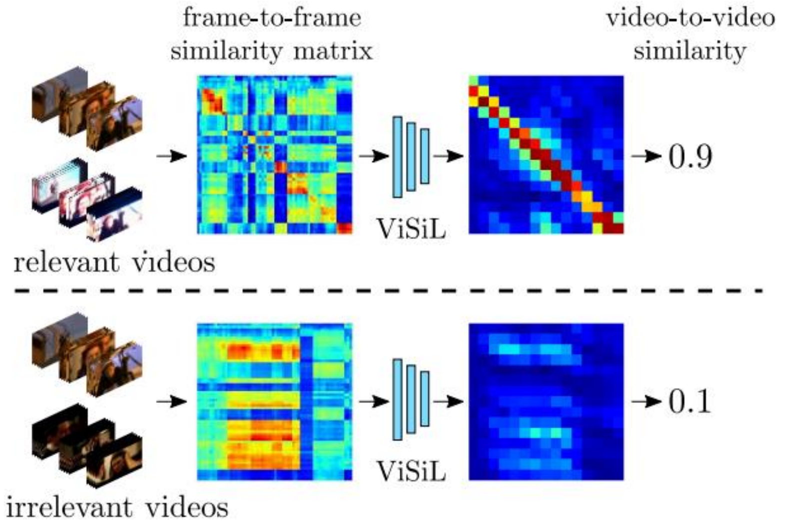


Figure 1. Depiction of the frame-to-frame similarity matrix and the CNN output of the ViSiL approach for two video pair examples: relevant videos that contain footage from the same incident (top), unrelated videos with spurious visual similarities (bottom).

Overview

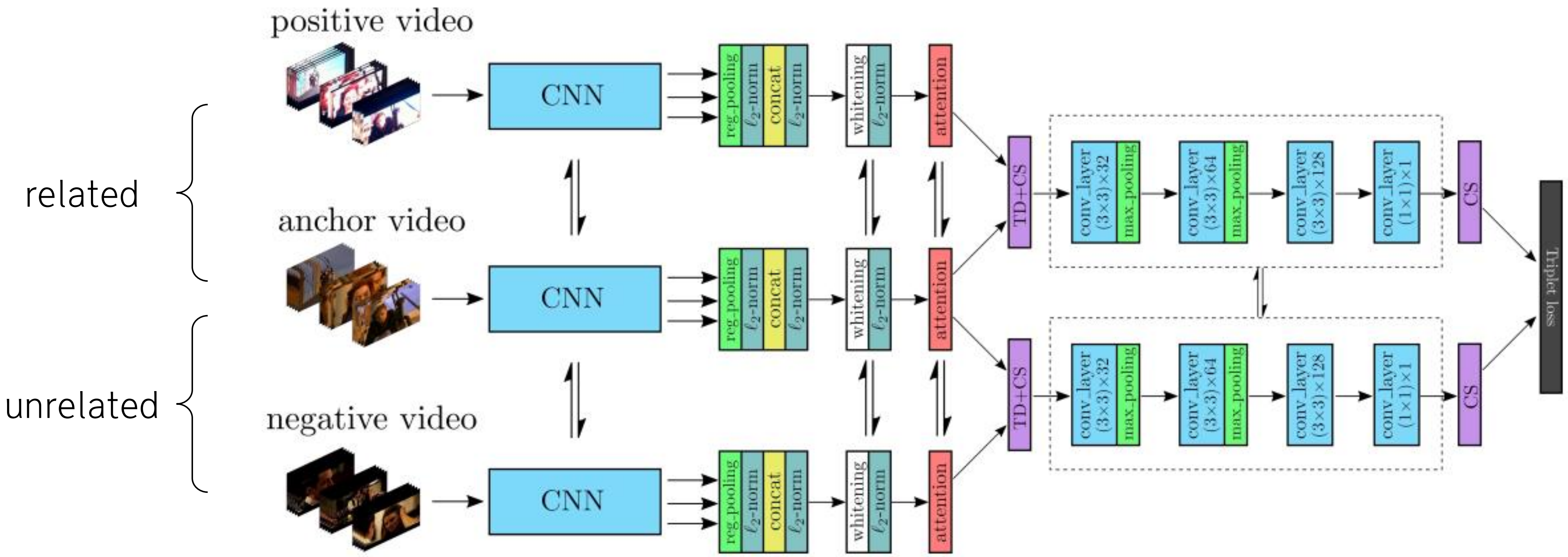


Figure 2. Overview of the training scheme of the proposed architecture. A triplet of an anchor, positive and negative videos is provided to a CNN to extract regional features that are PCA whitened and weighted based on an attention mechanism. Then the Tensor Dot product is calculated for the anchor-positive and anchor-negative pairs followed by Chamfer Similarity to generate frame-to-frame similarity matrices. The output matrices are passed to a CNN to capture temporal relations between videos and calculate video-to-video similarity by applying Chamfer Similarity on the output. The network is trained with the triplet loss function. The double arrows indicate shared weights.

Preliminaries

Tensor Dot (TD)

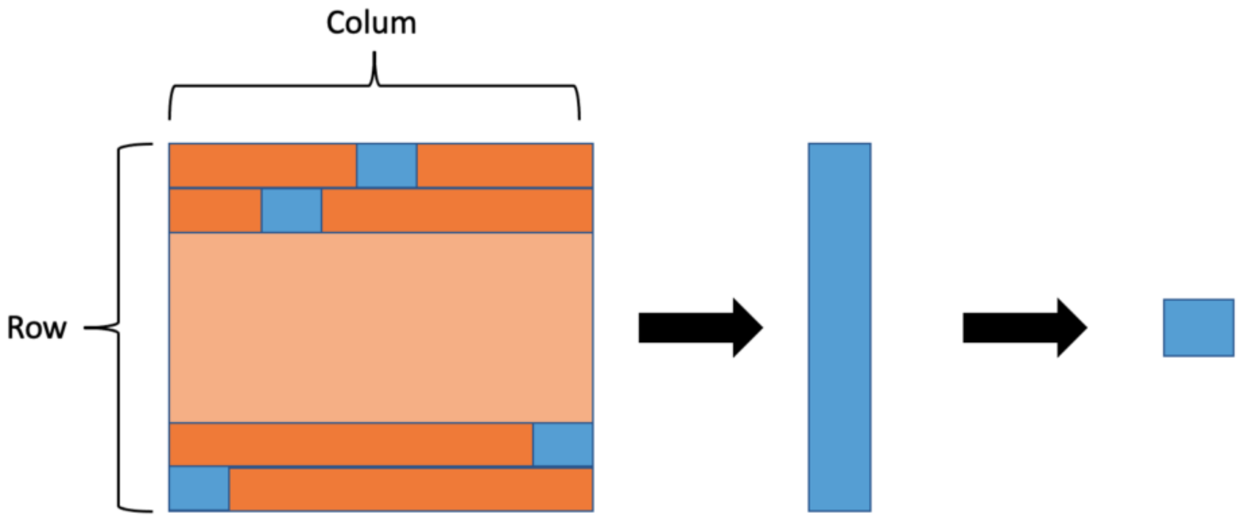
$$\mathcal{A} \in \mathbb{R}^{N_1 \times N_2 \times K}$$
$$\mathcal{B} \in \mathbb{R}^{K \times M_1 \times M_2}$$

$$\mathcal{C} = \mathcal{A} \cdot_{(i,j)} \mathcal{B}$$

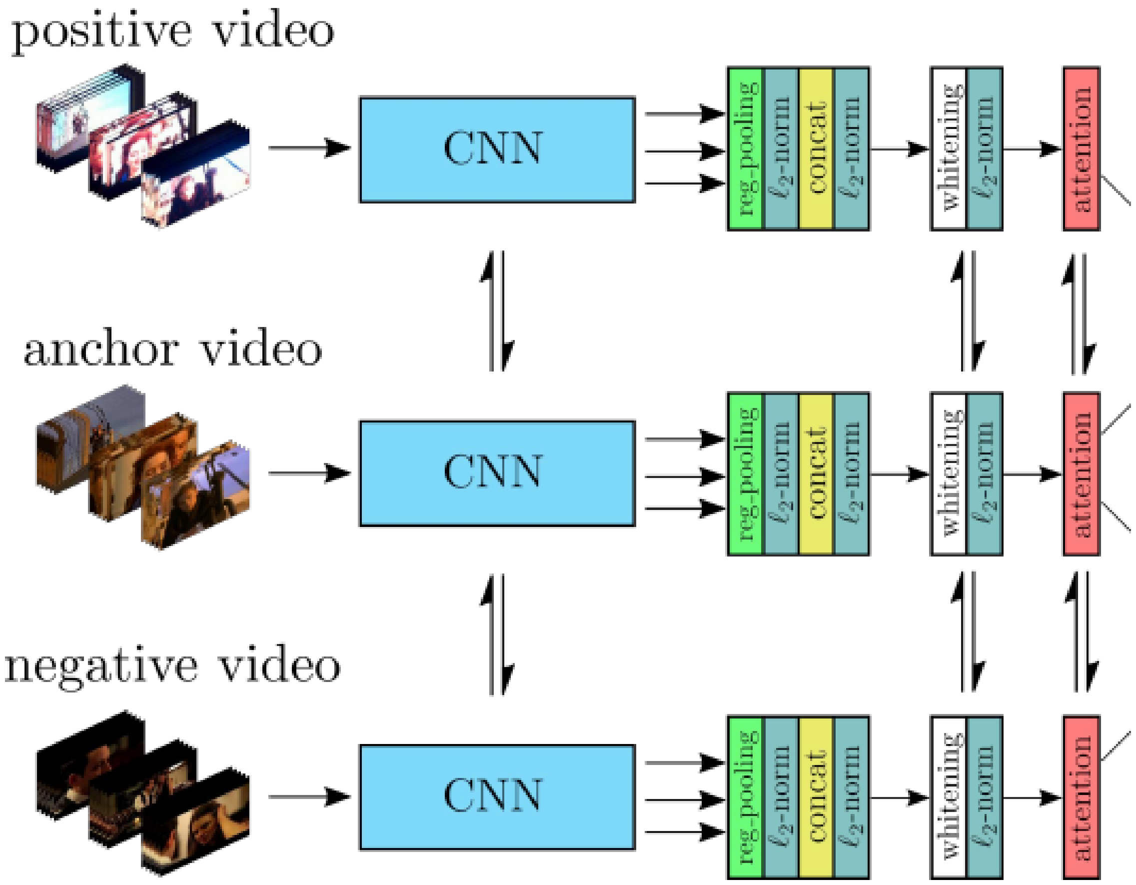
$$\mathcal{C} \in \mathbb{R}^{N_1 \times N_2 \times M_1 \times M_2}$$

Chamfer Similarity (CS)

$$CS(x, y) = \frac{1}{N} \sum_{i=1}^{N} \max_{j \in [1,M]} S(i, j)$$



ViSiL description : feature extraction



- Backbone : ResNet50
- feature maps from CNN
- R-MAC (Regional Maximum Activation of Convolution)

$$\mathcal{M}^k = \mathbb{R}^{N \times N \times C_k}$$

$$\mathcal{M} = \mathbb{R}^{N \times N \times C} \qquad C = C_1 + \ldots + C_K$$

- L2-normalization
- Concatenation
- L2-normalization

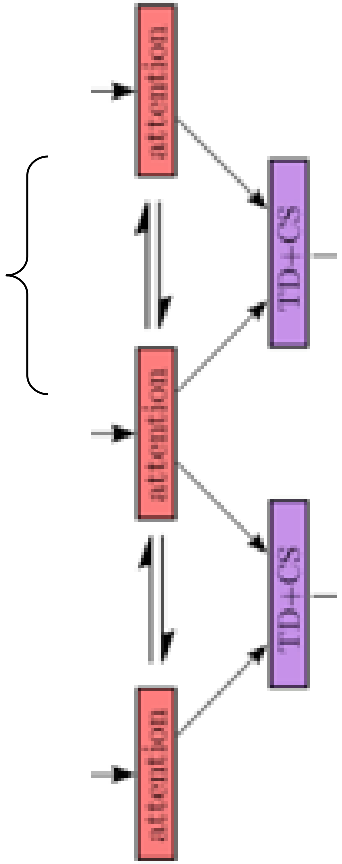
- Whitening (PCA)
- L2-normalization

- attention

$$\alpha_{ij} = \mathbf{u}^\top \mathbf{r}_{ij}, \quad s.t. \, \|\mathbf{u}\| = 1$$

$$\mathbf{r}'_{ij} = (\alpha_{ij}/2 + 0.5)\mathbf{r}_{ij}$$

ViSiL description : frame-to-frame similarity



Two video frames *b*, *d*

$$\mathcal{M}_d, \mathcal{M}_b = \mathbb{R}^{N \times N \times C}$$

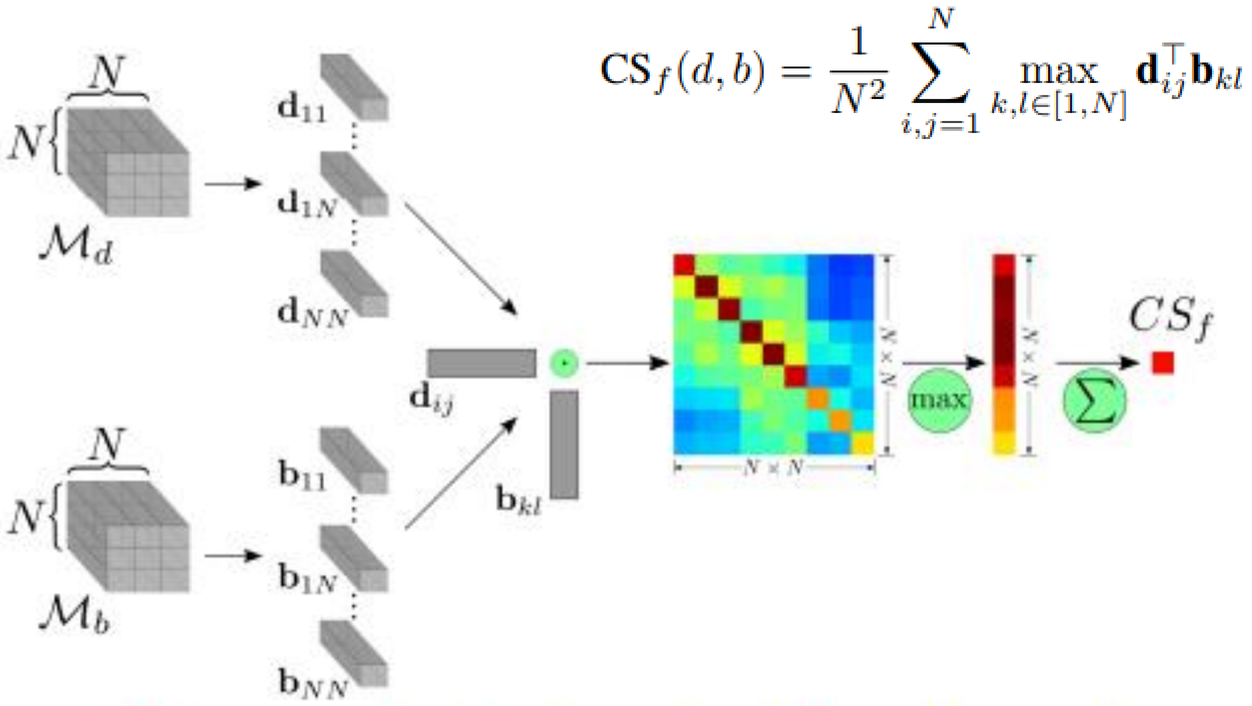
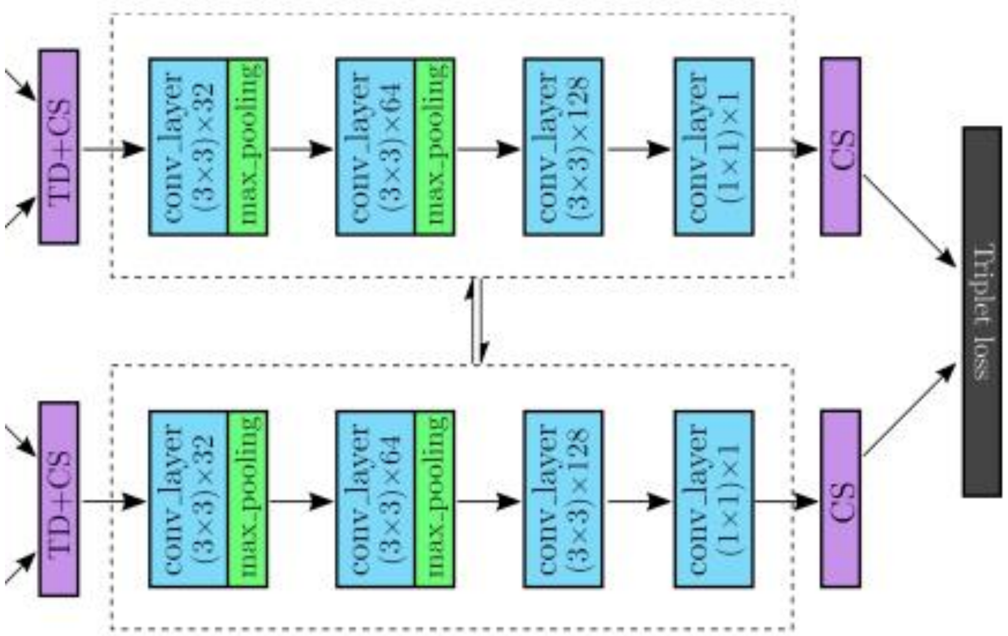


Figure 3. Illustration of frame-level similarity calculation between two video frames. In this example, the frames are near duplicates.

ViSiL description : video-to-video similarity

Two videos q, p with X and Y frames respectively.

$$\mathcal{S}_f^{qp} = \frac{1}{N^2}\sum_{i=1}^{N^2} \max_{j\in[1,N^2]} \mathcal{Q} \cdot_{(3,1)} \mathcal{P}^\top(\cdot, i, j, \cdot)$$



Type	**Kernel size / stride**	**Output size**	**Activ.**
Conv	3×3 / 1	$X \times Y \times 32$	ReLU
M-Pool	2×2 / 2	$X/2 \times Y/2 \times 32$	—
Conv	3×3 / 1	$X/2 \times Y/2 \times 64$	ReLU
M-Pool	2×2 / 2	$X/4 \times Y/4 \times 64$	—
Conv	3×3 / 1	$X/4 \times Y/4 \times 128$	ReLU
Conv	1×1 / 1	$X/4 \times Y/4 \times 1$	—

Table 1. Architecture of the proposed network for video similarity learning. For the calculation of the output size, we assume that two videos with total number of X and Y frames are provided.

$$\mathrm{CS}_v(q,p) = \frac{1}{X'}\sum_{i=1}^{X'} \max_{j\in[1,Y']} \mathrm{Htanh}(\mathcal{S}_v^{qp}(i,j)) \qquad \mathcal{S}_v^{qp} \in \mathbb{R}^{X'\times Y'}$$

ViSiL description : Loss function

Target video similarity score : $\mathbf{CS}_v(q,p)$ (v, v^+, v^-).

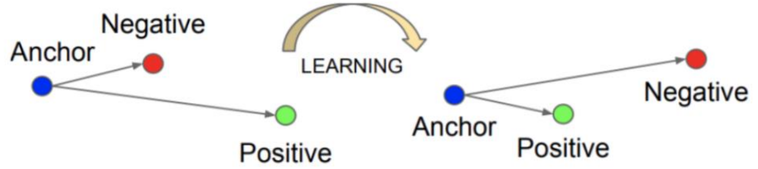


Figure 3. The **Triplet Loss** minimizes the distance between an *anchor* and a *positive*, both of which have the same identity, and maximizes the distance between the *anchor* and a *negative* of a different identity.

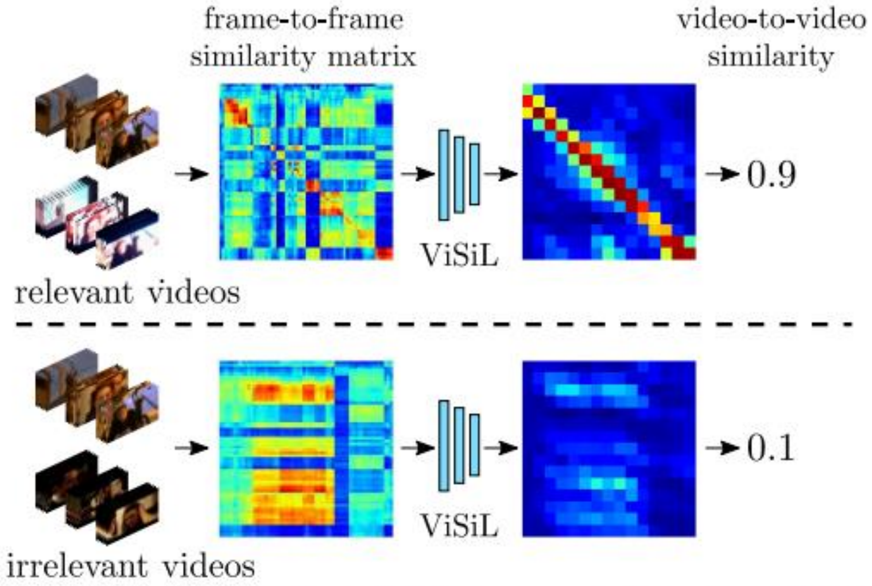


Figure 1. Depiction of the frame-to-frame similarity matrix and the CNN output of the ViSiL approach for two video pair examples: relevant videos that contain footage from the same incident (top), unrelated videos with spurious visual similarities (bottom).

$$\mathcal{L}_{tr} = \max\{0, \mathbf{CS}_v(v, v^-) - \mathbf{CS}_v(v, v^+) + \gamma\}$$

$$\mathcal{L}_{reg} = \sum_{i=1}^{X'}\sum_{j=1}^{Y'} |\max\{0, \mathcal{S}_v^{qp}(i,j) - 1\}| + \\ + |\min\{0, \mathcal{S}_v^{qp}(i,j) + 1\}|$$

$$\mathcal{L} = \mathcal{L}_{tr} + r * \mathcal{L}_{reg}$$

Evaluation

The proposed approach is evaluated on four retrieval tasks, reporting mean Average Precision (mAP)

- NDVR (Near-Duplicate Video Retrieval)
- FIVR (Fine-grained Incident Video Retrieval)
- EVR (Event Video Retrieval)
- AVR (Action Video Retrieval)

Evaluation : Datasets

VCDB : used as training datasets, in order to make triplets.

CC_WEB_VIDEO : NDVR

FIVR-5K : FIVR

EVVE : EVR

ActivityNet : AVR

Experiments : frame-to-frame similarity

Features	Dims.	DSVR	CSVR	ISVR
MAC [33]	2048	0.747	0.730	0.684
SPoC [1]	2048	0.735	0.722	0.669
R-MAC [33]	2048	0.777	0.764	0.707
GeM [12]	2048	0.776	0.768	0.711
iMAC [20]	3840	0.755	0.749	0.689
L_2-iMAC	4x3840	0.814	0.810	0.738
L_2-iMAC	4x512	0.804	0.802	0.727
L_3-iMAC	9x3840	**0.838**	**0.832**	**0.739**
L_3-iMAC	9x256	0.823	0.818	0.738

Table 2. mAP comparison of proposed feature extraction and similarity calculation against state-of-the-art feature descriptors with dot product for similarity calculation on FIVR-5K. Video similarity is computed based on CS on the derived similarity matrix.

Experiments : Ablation study

Task	DSVR	CSVR	ISVR
ViSiL$_f$	0.838	0.832	0.739
ViSiL$_f$**+W**	0.844	0.837	0.750
ViSiL$_f$**+W+A**	0.856	0.848	0.768
ViSiL$_{sym}$	0.830	0.823	0.731
ViSiL$_v$	**0.880**	**0.869**	**0.777**

Table 3. Ablation studies on FIVR-5K. **W** and **A** stand for whitening and attention mechanism respectively.

$\mathcal{L}_{reg}$	DSVR	CSVR	ISVR
✗	0.859	0.842	0.756
✓	**0.880**	**0.869**	**0.777**

Table 4. Impact of similarity regularization on the performance of the proposed method on FIVR-5K.

Experiments : comparison with state-of-the-art

Method	**cc_web**	**cc_web***	**cc_web$_c$**	**cc_web$_c^*$**
DML [21]	0.971	0.941	0.979	0.959
CTE [28]	**0.996**	—	—	—
DP [7]	0.975	0.958	0.990	0.982
TN [32]	0.978	0.965	0.991	0.987
ViSiL$_f$	0.984	0.969	0.993	0.987
ViSiL$_{sym}$	0.982	0.969	0.991	0.988
ViSiL$_v$	0.985	**0.971**	**0.996**	**0.993**

Table 5. mAP of three ViSiL setups and SoA methods on four different versions of CC_WEB_VIDEO. (*) denotes evaluation on the entire dataset, and subscript c that the cleaned version of the annotations was used.

Run	**DSVR**	**CSVR**	**ISVR**
LBoW [20]	0.710	0.675	0.572
DP [7]	0.775	0.740	0.632
TN [32]	0.724	0.699	0.589
ViSiL$_f$	0.843	0.797	0.660
ViSiL$_{sym}$	0.833	0.792	0.654
ViSiL$_v$	**0.892**	**0.841**	**0.702**

Table 6. mAP comparison of three ViSiL setups and state-of-the-art methods on the three tasks of FIVR-200K.

Experiments : comparison with state-of-the-art

Method	**mAP**	**per event class**												
LAMV[2]	0.536	0.715	0.383	0.158	0.461	0.387	0.277	0.247	0.138	0.222	0.273	0.273	0.908	0.691
LAMV+QE [2]	0.587	0.837	0.500	0.126	**0.588**	**0.455**	0.343	0.267	0.142	0.230	0.293	0.216	**0.950**	0.776
ViSiL$_f$	0.589	0.889	0.570	0.169	0.432	0.345	0.393	0.297	0.181	0.479	0.564	0.369	0.885	0.799
ViSiL$_{sym}$	0.610	0.864	0.704	**0.357**	0.440	0.363	0.295	**0.370**	0.214	0.577	0.389	0.266	0.943	0.702
ViSiL$_v$	**0.631**	**0.918**	**0.724**	0.227	0.446	0.390	**0.405**	0.308	**0.223**	**0.604**	**0.578**	**0.399**	0.916	**0.855**

Table 7. mAP comparison of three ViSiL setups with the LAMV [2] on EVVE. The ordering of events is the same as in [28]. Our results are reported on a subset of the videos (≈80% of the original dataset) due to unavailability of the full original dataset.

Method	mAP
DML [21]	0.705
VReL [10]	0.209
DP [7]	0.621
TN [32]	0.648

Method	mAP
ViSiL$_f$	0.652
ViSiL$_{sym}$	**0.745**
ViSiL$_v$	0.710

Table 8. mAP comparison of three ViSiL setups and four publicly available retrieval methods on ActivityNet based on the reorganization from [10].

Conclusion

ViSiL : a network that learns to compute **similarity between pairs of videos**

Key contributions

a) a **frame-to-frame similarity** computation scheme that **captures similarities at regional level**

b) a **supervised video-to-video similarity computation scheme** that analyzes the **frame-to-frame similarity matrix** to robustly establish high similarities between video segments of the compared videos.

video similarity computation method
that is accounting for both the **fine-grained spatial and temporal aspects of video similarity**

감사합니다 👍