RELEVANCE-AWARE ONLINE MINING FOR VIDEO RETRIEVAL

Falcon A., Serra G., Lanz O. - Univ. of Udine & FBK & Free Univ. of Bozen
falcon.alex@spes.uniud.it, giuseppe.serra@uniud.it, lanz@inf.unibz.it

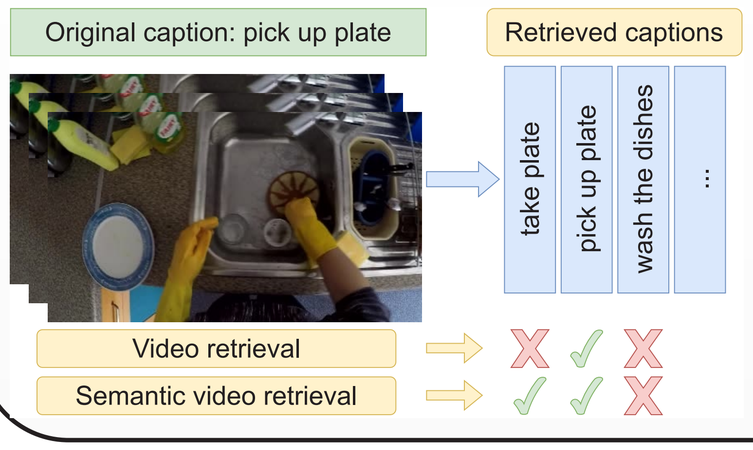
ICVSS 2022
Sicily ~ 10-16 July
International Computer Vision Summer School

Abstract

Semantic text-video retrieval (SVR) is a recently introduced problem where the quality of the full ranking lists is used to assess the performance. To tackle video retrieval, only video-caption pairs in the dataset are considered as relevant to each other. This approach does not transfer well to SVR. We propose RANP to identify new video-caption pairs by using the relevance to separate irrelevant from relevant content. With RANP, considerable improvements are observed on two public datasets.

Semantic video retrieval

Instance-based → mean rank, R@K, *etc*
Semantic → nDCG, mAP.



Relevance

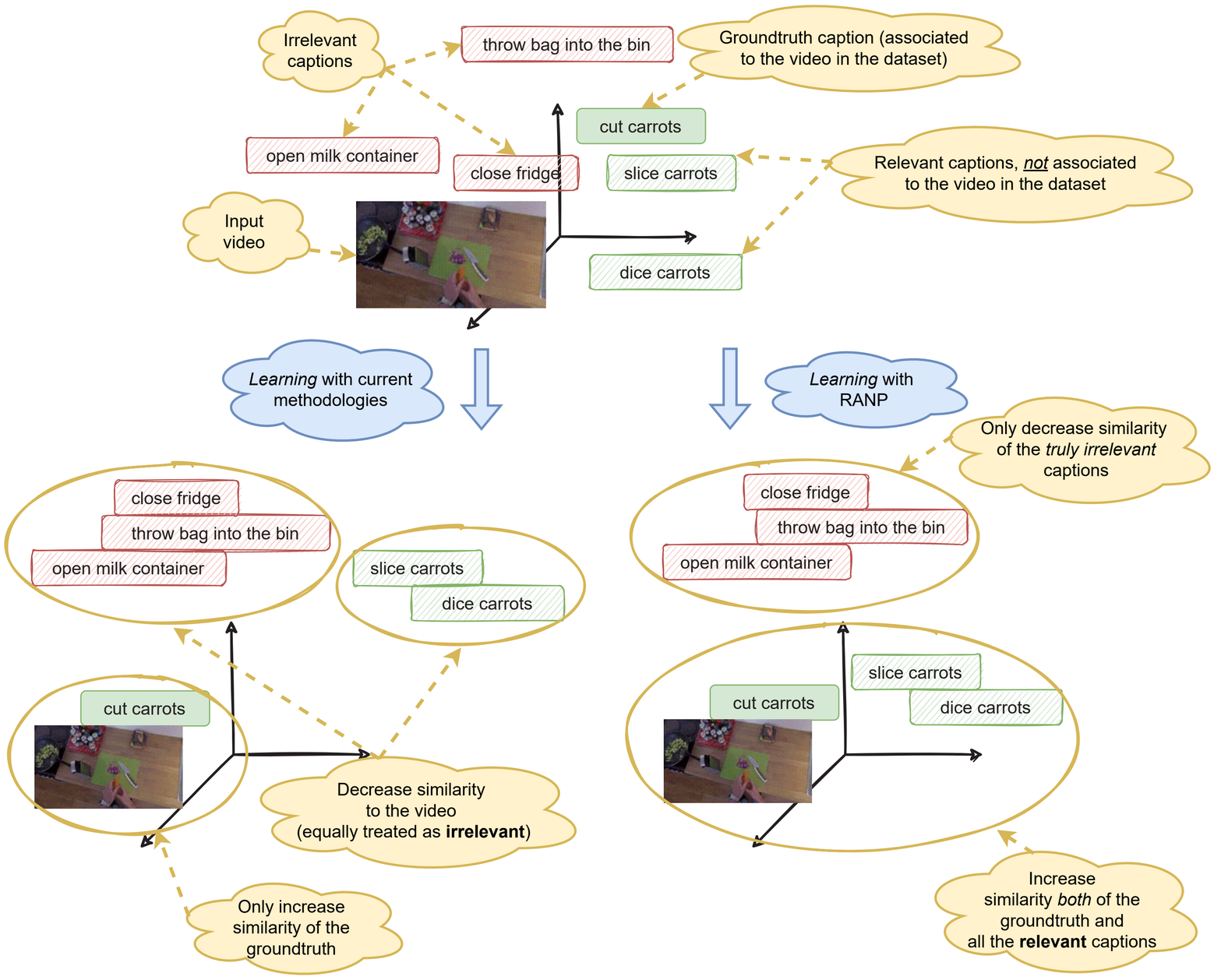
The **relevance** quantifies the degree of "closeness" of two input items. We consider a relevance function $\mathcal{R}$ [3] defined on noun and verb classes. Therefore, two captions (or videos) have a high relevance if they share similar verbs and nouns (synonyms included). *E.g.*

$\mathcal{R}$(take plate , pick up plate) = 1

$\mathcal{R}$(pick up plate , grab knife) =0.5

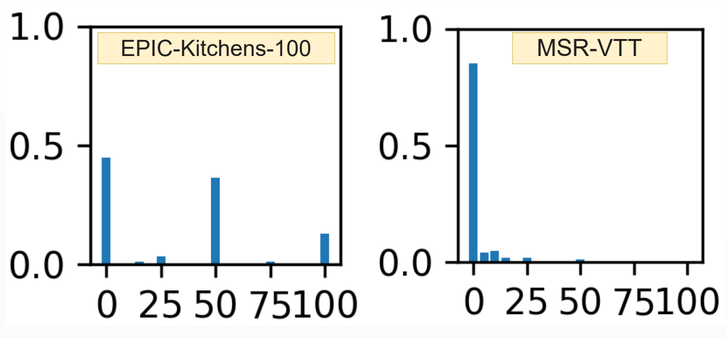
Relevance-Aware Negative and Positives mining (RANP)

At training time, we leverage $\mathcal{R}$ to identify which captions are **relevant** to the input video and separate them from the **irrelevant** examples. Then, we increase the similarity of all the multiple relevant captions, while decreasing the similarity to irrelevant ones.



Distribution of relevance

How many captions are treated as irrelevant, although they are actually relevant to the video?



Quantitative results

On **EPIC-Kitchens-100**, compared to:

- baseline: +23% nDCG, +8% mAP;
- SoTA: +5% nDCG, +3% mAP.

On **MSR-VTT**, compared to:

- baseline: +6% nDCG;
- SoTA: +2% nDCG.

References

[1] A., Falcon, G., Serra, O., Lanz, Learning video retrieval models with relevance-aware online mining, in *ICIAP*, 2022

[2] M., Wray, H., Doughty, D., Damen, On Semantic Similarity in Video Retrieval, in *CVPR*, 2021

[3] D., Damen, et al., Rescaling Egocentric Vision, in *IJCV*, 2021

Acknowledgements

We gratefully acknowledge the support from Amazon AWS Machine Learning Research Awards (MLRA) and NVIDIA AI Technology Centre (NVAITC), EMEA. We acknowledge the CINECA award under the ISCRA initiative, which provided computing resources for this work.

Code at: