Semantic Role Aware Correlation Transformer For Text To Video Retrieval




Burak Satar[1,2] Hongyuan Zhu[1] Xavier Bresson[3] Joo Hwee Lim[1,2]

[1]Institute for Infocomm Research, A*STAR

Agency for Science, Technology and Research SINGAPORE

[2]School of Computer Science and Engineering, NTU

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

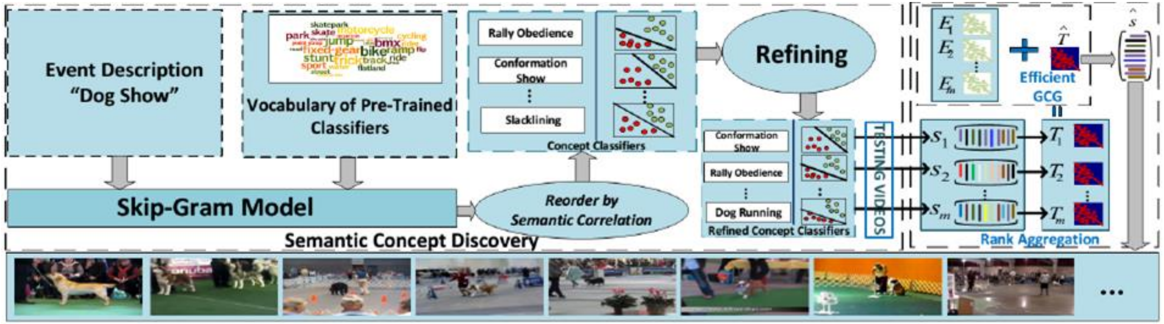
[3]Department of Computer Science, NUS

NUS National University of Singapore

Research Problem

- The amount of video available online is increasing.
  - 34K hours of video upload every day at Youtube
  - Surveillance cameras, car cameras, personal cameras

- Conventional models* are based on keywords query.
  - Limited and insufficient to retrieve fine-grained and compositional events.

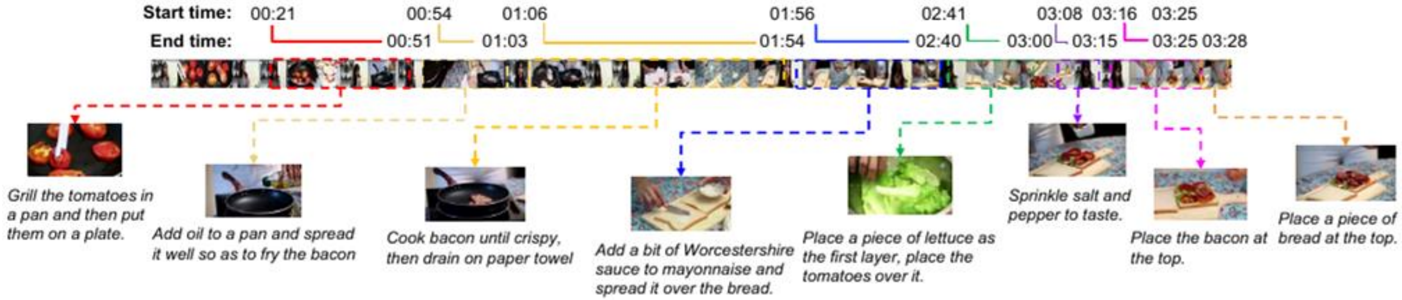


* [Semantic Concept Discovery for Large-Scale Zero-Shot Event Detection](), Chang et al., IJCAI'15
* [Composite Concept Discovery for Zero-Shot Video Event Detection](), Habibian et al., ICMR'14

Task: Text to Video Retrieval

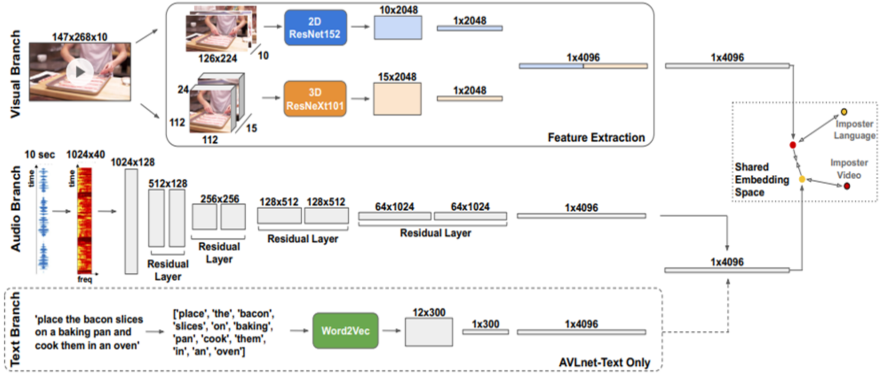
- Given a textual query, ranking all the video candidates such that the video associated with the textual query is ranked as high as possible.

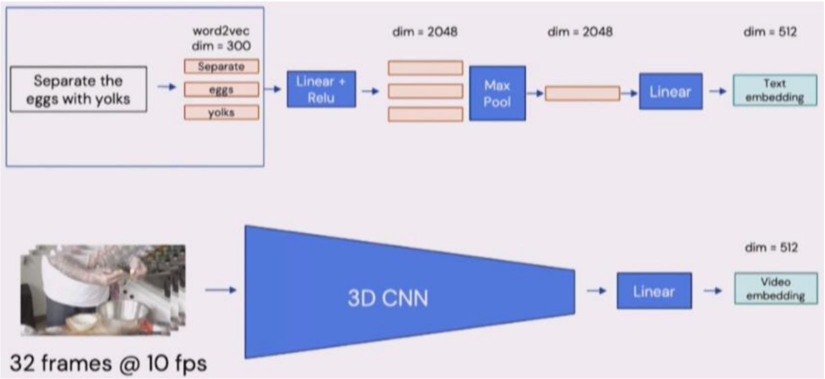
Dataset: YouCook2



- While the training set includes ~9.5k clips, the validation set has ~3.3k clips.
- Validation set is used for evaluation since the test set has no annotations.

Towards Automatic Learning of Procedures From Web Instructional Videos, Zhou et al., AAAI'18

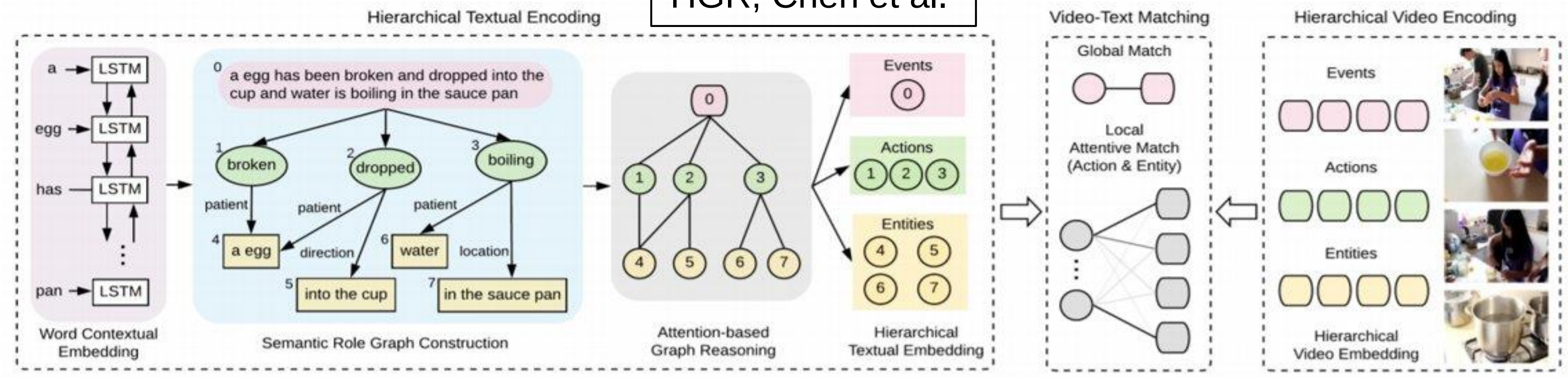




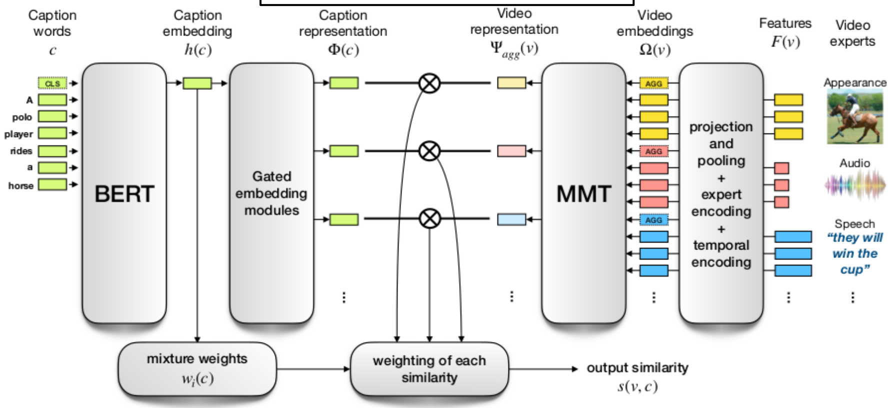
- Having only one joint embedding space causes losing fine-grained details.
- Some other papers try both global and local; however, still a semantic gap.

End-to-End Learning of Visual Representations from Uncurated Instructional Videos, Miech et al., CVPR'20
AVLnet: Learning Audio-Visual Language Representations from Instructional Videos, Rouditchenko et al., arXiv'20

HGR, Chen et al.

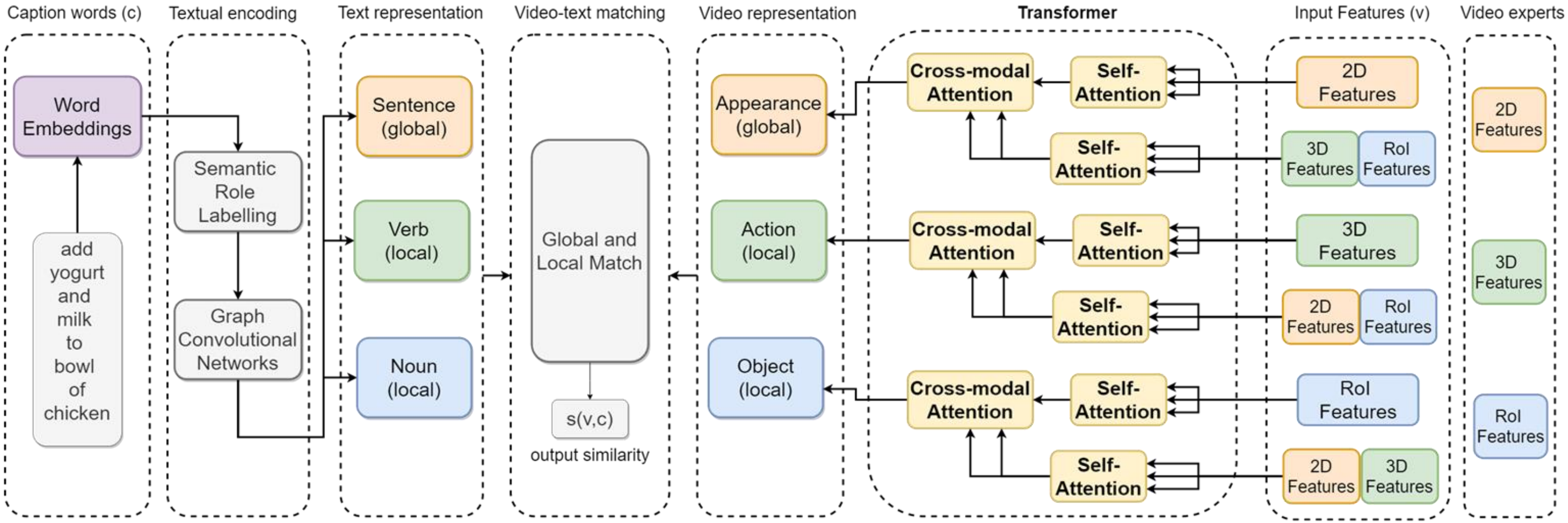


MMT, Gabeur et al.



Fine-grained Video-Text Retrieval with Hierarchical Graph Reasoning, Chen et al., CVPR'20
MMT: Multi-modal Transformer for Video Retrieval, Gabeur et al., ECCV'20

Model



Task: Text-to-video retrieval

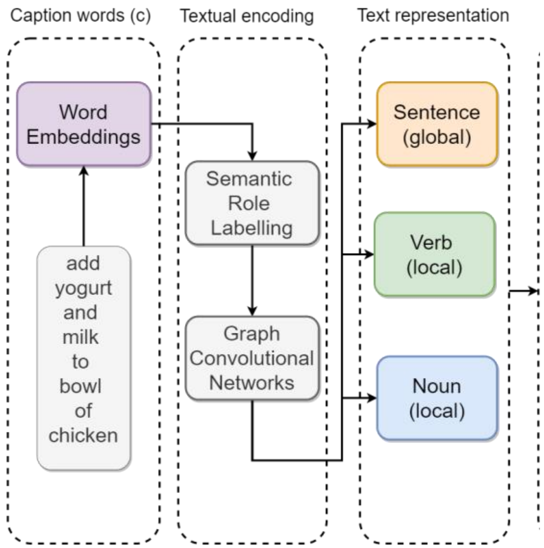
Dataset: YouCook2

Textual Encoding Part

Semantic Role Labelling


```
"cut the roll with a sharp knife":
[
    {
        "ROOT": {
            "words": ["cut", "the", "roll", "with", "a", "sharp", "knife"],
            "spans": [0, 1, 2, 3, 4, 5, 6],
            "role": "ROOT"
        },
        "1": {
            "role": "V",
            "spans": [0],
            "words": ["cut"]
        },
        "2": {
            "role": "ARG1",
            "spans": [1, 2],
            "words": ["the", "roll"]
        },
        "3": {
            "role": "ARGM-MNR",
            "spans": [3, 4, 5, 6],
            "words": ["with", "a", "sharp", "knife"]
        }
    },
    [
        ["1", "2", "ARG1"],
        ["1", "3", "ARGM-MNR"]
    ]
],
```


https://demo.allennlp.org/semantic-role-labeling



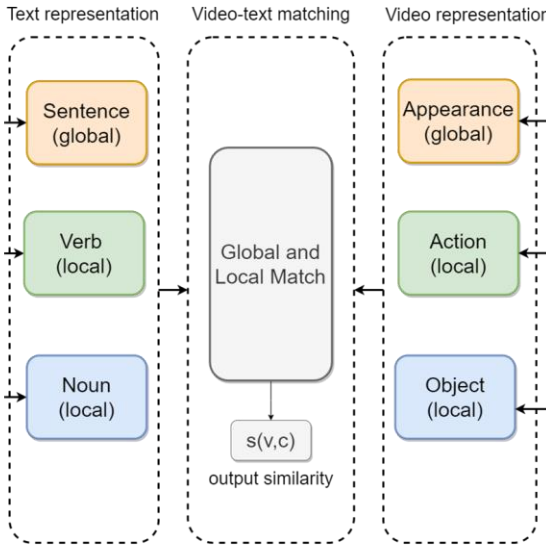
Text-Video Matching Part

- Cosine similarity score for each level.

$$s(V,C) = \frac{< v, c >}{||v||_2||c||_2}$$

- We average similarities and utilize contrastive ranking loss as a training objective.

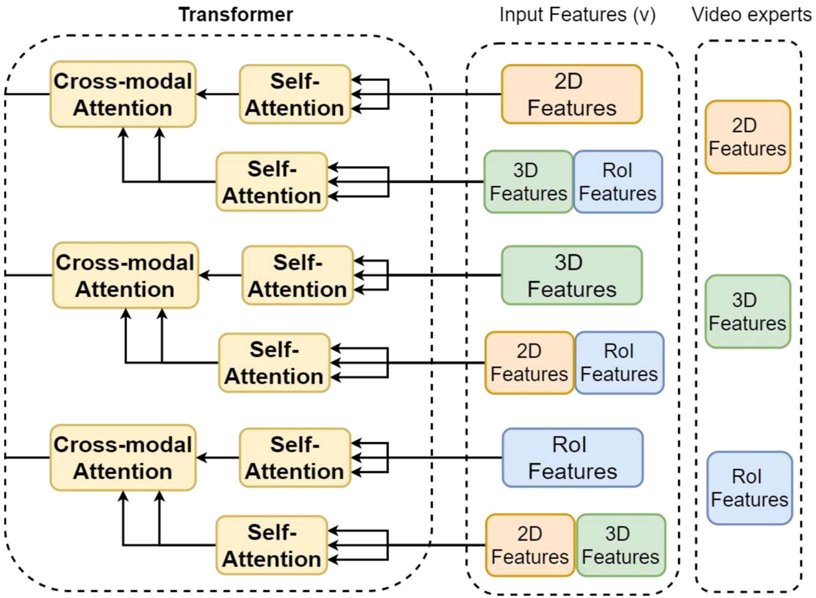
$$L(v_p, c_p) = [\Delta + s(v_p, c_n) - s(v_p, c_p)]+ \\ [\Delta + s(v_n, c_p) - s(v_p, c_p)]$$



Visual Encoding Part

$$f_e = \text{Concat}(F_T, F_O)$$
$$z_e = \text{Norm}\big(\text{MultiHead}(f_e, f_e, f_e) + f_e\big)$$
$$s_e = \text{Norm}\big(\text{FF}(z_e) + z_e\big)$$

$$z_s = \text{Norm}\big(\text{MultiHead}(F_S, F_S, F_S) + F_S\big)$$
$$c_e = \text{Norm}\big(\text{MultiHead}(z_s\ ,\ s_e\ ,\ s_e\ ) + z_s\big)$$
$$E_S = \text{Norm}\big(\text{FF}(c_e) + c_e\big)$$



Result

Method	Pre-training	Visual Backbone	Batch Size	R@1↑	R@5↑	R@10↑	MedR↓
Random	No	-	-	0.03	0.15	0.3	1675
Miech et al [6]	No	ResNeXt-101	-	4.2	13.7	21.5	65
HGLMM [28]	No	-	-	4.6	14.3	21.6	75
HGR [3]	No	ResNeXt-101	32	4.7	14.1	20.0	87
Ours	No	ResNeXt-101	32	**5.3**	**14.5**	20.8	77
Miech et al+FT [6]	HowTo100M	ResNeXt-101	-	8.2	24.5	35.3	24
ActBert [17]	HowTo100M	ResNet-3D	-	9.6	26.7	38.0	19
MMV FAC [18]	HowTo100M+AudioSet	TSM-50	4096	11.5	30.2	41.5	16
MIL-NCE [7]	HowTo100M	S3D	8192	15.1	38.0	51.2	10

- Text-to-video retrieval comparison with SOTA approaches on YouCook2 validation set.
- Our method surpasses the SOTA methods in the first two parameters without pre-training.

Ablation

Method	Visual Features			Feature Dimension	R@1↑	R@5↑	R@10↑	MedR↓
	Appearance	Action	Object					
HGR [3] : Ours	2D	2D	2D	2048	4.7 : 4.2	13.8 : 13.7	19.7 : 19.4	86 : 86
HGR [3] : Ours	2D + 3D	2D + 3D	2D + 3D	2048	4.8 : 4.5	14.0 : 13.2	20.3 : 20.0	85 : 85
HGR [3] : Ours	2D + 3D	2D + 3D	2D + 3D	4096	4.8 : 4.5	14.0 : 13.2	20.3 : 20.0	85 : 85
HGR [3] : Ours	2D	3D	RoI	2048	4.7 : **5.3**	14.1 : **14.5**	20.0 : **20.8**	87 : **77**

- Ablation studies to investigate the contributions of various feature experts at different levels.
- This confirms our insight that inter-modal correlation can be exploited with our proposed cross-modal attention mechanism to achieve better results.

Summary

- Our model surpasses a strong baseline with a high margin in all metrics.
- It also overpasses other SOTA methods in R@1, R@5 metrics.
- We think that modality-specific and modality-complement features improve accuracy at R@1 and R@5, which are more demanding and useful for real-world applications.

Thank you for watching!

https://buraksatar.github.io/

Burak Satar

Hongyuan Zhu

Xavier Bresson

Joo Hwee Lim

Agency for Science, Technology and Research SINGAPORE

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

NUS National University of Singapore