Semantic Role Aware Correlation Transformer for Text to Video Retrieval

Burak Satar[1,2], Zhu Hongyuan[1,4], Xavier Bresson[3], Joo Hwee Lim[1,2]

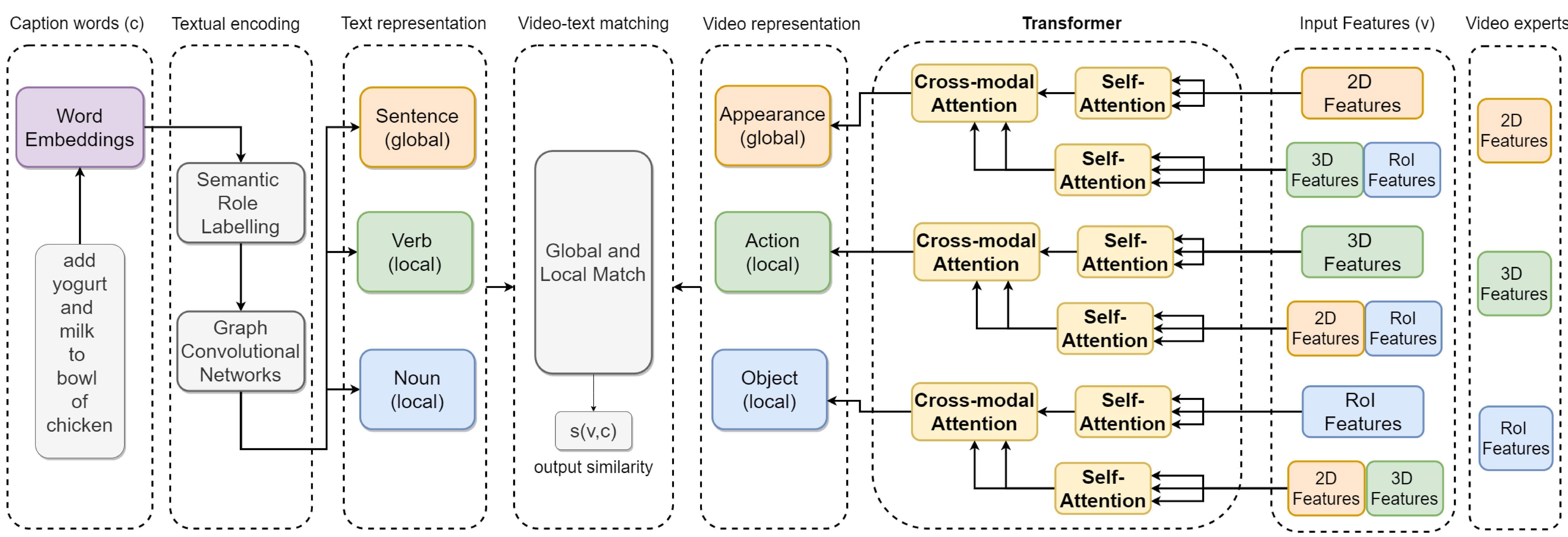
Overview

Introduction. Retrieving related video on a textual query gets harder since the number of videos on the internet increases. Most works use one joint embedding space for text-to-video retrieval task without fully exploiting cross-modal features. We propose a hierarchical model representing complex textual and visual features with three joint embedding spaces by utilizing self-attention and cross-modal attention to exploit the modality-specific and modality-complement visual embeddings. Preliminary results show that our approach surpasses a current state-of-the-art method, with a high margin in all metrics. It also overpasses two SOTA methods in terms of two metrics.

Related Work. Conventional models are based on keywords query, which is insufficient to retrieve fine-grained and compositional events [1]. Most works [4] embeds whole videos and texts into flat vectors to exploit global features, while the others focus on only local features. Recently, Chen et al. [3] propose decomposing text into three semantic roles (events, actions and entities) and then embedding 2D video features into these three spaces accordingly for matching. Another line of research uses a BERT-like transformer [5] to learn the text-video correspondence, based on recent mixture-of-expert embedding [6], which requires a large-scale dataset for pre-training.

Experiments. We evaluate our model on YouCook2 [2], which is a video dataset on cooking gathered from YouTube. The task is retrieving video clips based on text queries. R@1, R@5, R@10 and median rank are used as evaluation metrics.

Method



We propose a novel transformer architecture for video-text matching inspired by [3] and [5]. Different from [3], which only considers multi-head embedding of the spatial frame and ignores the interaction between different visual contexts, our method explicitly considers more fine-grained visual encoding of object, spatial and temporal contexts by embedding RoI regions, 2D frames and video sequences into the corresponding space with their interactions. Different from [5], which only uses self-attention to discover modality-specific information, our method uses a self-attention scheme to discover modality-specific discriminative features. Also, our model utilizes cross-modal attention to consider the interactions between object, spatial and temporal contexts to discover modality-complement features for better align video and text.

Visual & Textual Encoding

Textual encoding. We follow Chen et al. [3] to disentangle text embeddings. First, semantic role labelling, then GCN is applied.
Visual encoding. Our aim is to calculate final embeddings for each level. This formula only shows our implementation on the spatial level.

$$f_e = \text{Concat}(F_T, F_O)$$
$$z_e = \text{Norm}(\text{MultiHead}(f_e, f_e, f_e) + f_e)$$
$$s_e = \text{Norm}(\text{FF}(z_e) + z_e)$$
$$z_s = \text{Norm}(\text{MultiHead}(F_S, F_S, F_S) + F_S)$$
$$c_e = \text{Norm}(\text{MultiHead}(s_e, z_s, z_s) + z_s)$$
$$E_S = \text{Norm}(\text{FF}(c_e) + c_e)$$

Cross-modal matching. We utilize cosine similarity to calculate the score for each level by corresponding visual & textual embeddings. We average the similarities and utilize contrastive ranking loss as a training objective.

Results

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

Table 1. Text-to-video retrieval comparison with SOTA approaches on YouCook2 validation set. Our method surpasses the SOTA methods in the first two parameters when without pre-training.

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**

Table 2. 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.

Conclusion. Our model surpasses a strong baseline with a high margin, and 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.

Authors

[1] Institute for Infocomm Research, A*STAR, Singapore
[2] School of Computer Science and Engineering, NTU, Singapore
[3] Department of Computer Science, National University of Singapore
[4] Corresponding Author
{burak_satar, zhuh, joohwee}@i2r.a-star.edu.sg, xavier@nus.edu.sg

Acknowledgements

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project A18A2b0046). Xavier Bresson is supported by the NRF Fellowship NRFF2017-10.

References

[1] X. Chang, et al., "Semantic concept discovery for large-scale zero-shot event detection," in IJCAI, 2015
[2] L. Zhou, C. Xu, and J. Corso, "Towards automatic learning of procedures from web instructional videos," in AAAI, 2018, pp. 7590–7598.
[3] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in CVPR, 2020
[4] Y. Liu, et al., "Use what you have: Video retrieval using representations from collaborative experts," in arXiv, 2019.
[5] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal Transformer for Video Retrieval," in ECCV, 2020
[6] A. Miech, et al., "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips," in ICCV, 2019