Hierarchical Semantic Correspondence Networks for Video Paragraph Grounding

Chaolei Tan[1] Zihang Lin[1] Jian-Fang Hu[123*] Wei-Shi Zheng[123] Jianhuang Lai[123]

[1]School of Computer Science and Engineering, Sun Yat-sen University, China

[2]Guangdong Province Key Laboratory of Information Security Technology, China

[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

Poster Session THU-AM-238

Highlights

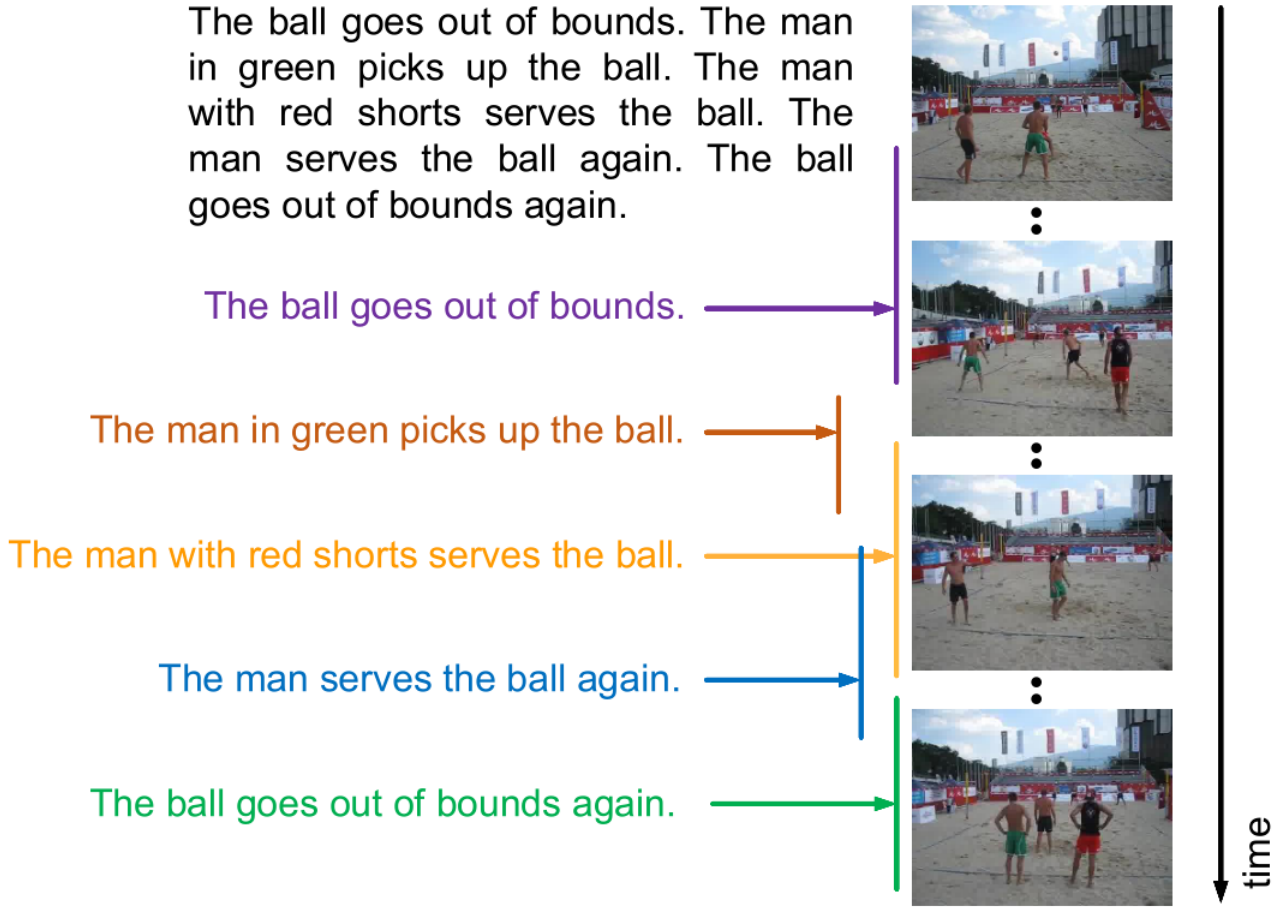
- We investigate the problem of how to explore hierarchical visual-textual correspondence and ground multiple levels of queries for Video Paragraph Grounding (VPG).
- Motivated by the inherent hierarchies of videos and paragraphs, we naturally introduce a pioneering hierarchical modeling framework without additional external components for VPG.
- We design a novel encoder-decoder network to explicitly learn and exploit the hierarchical cross-modal alignment and dense multi-level supervision.
- The model achieves new state-of-the-art performance and shows a series of hierarchical perception capabilities for the cross-modal semantic association.

Backgrounds

- **Task**: Video Paragraph Grounding (VPG)

- **Definition**: Localize the timestamps of multiple visual events from an untrimmed video at once, where each of the visual events semantically corresponds to one of the sentences in the given paragraph description.

- To well address this problem, the key challenge lies on how to understand the content of a complicated paragraph and an untrimmed video as well as their complex semantic relations.

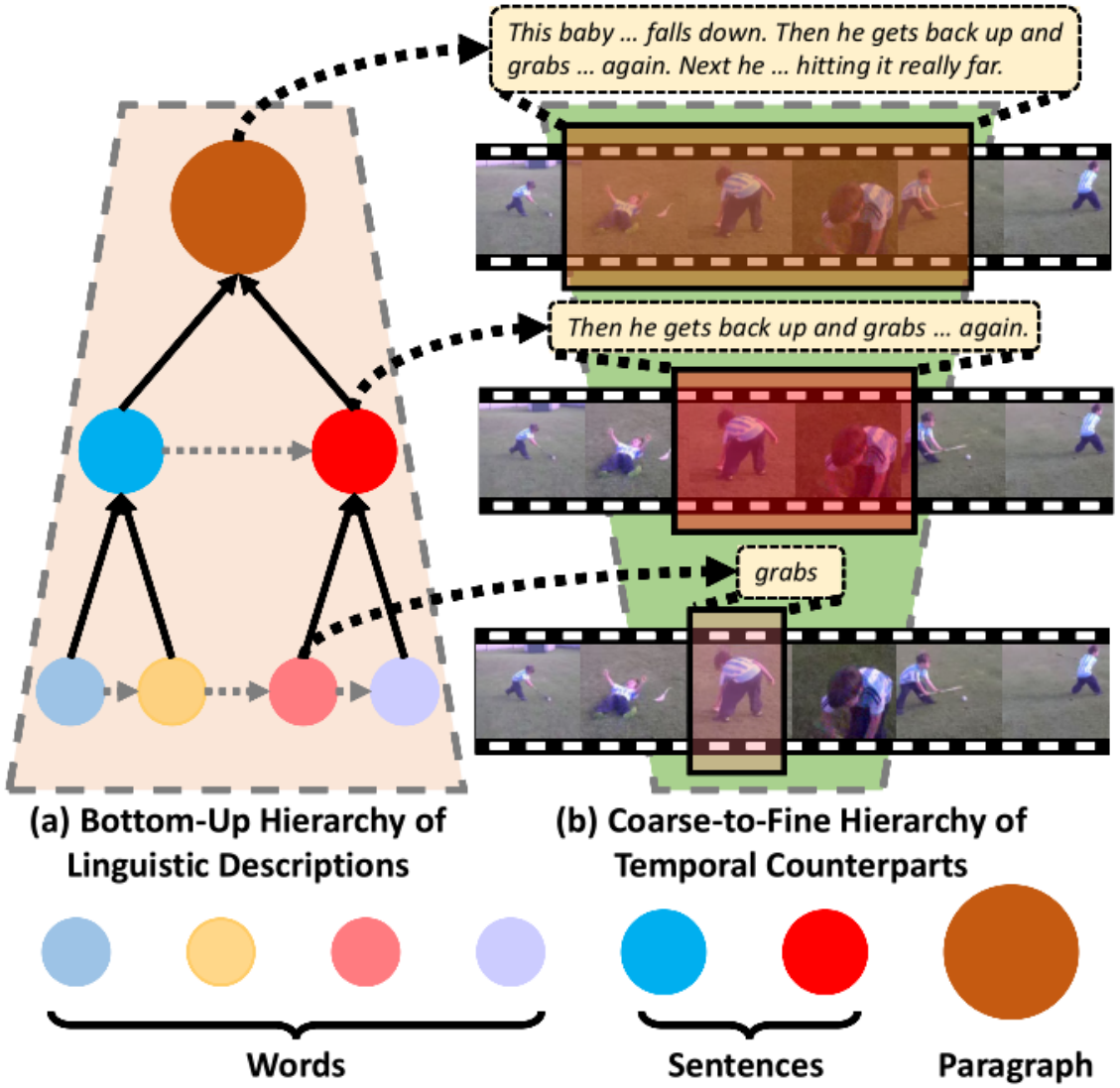


Motivation

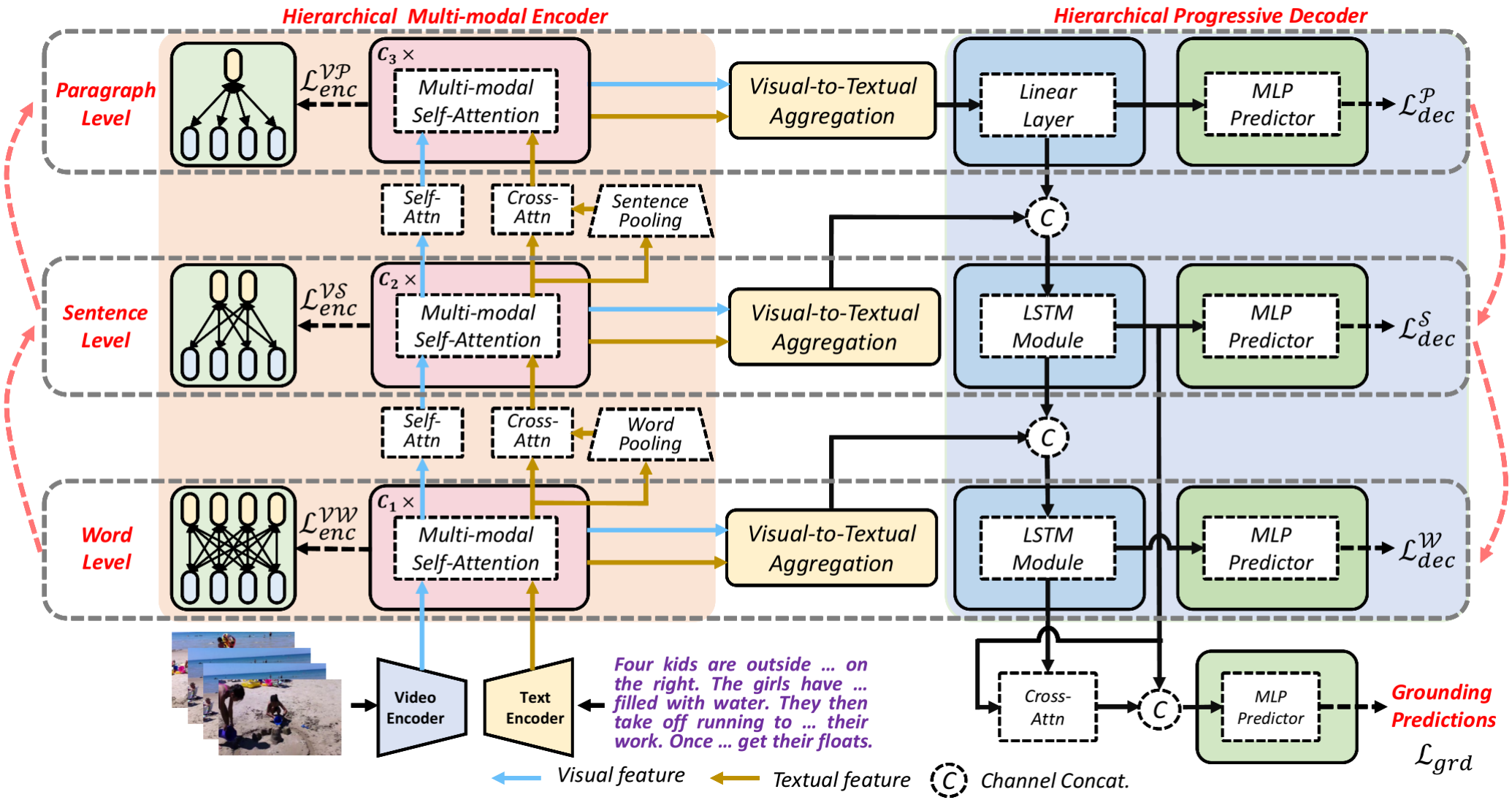
- Prior works model the video-text relations in VPG from a single sentence level. They neglect the rich cross-modal relations at other semantic levels, hence failing to benefit from the hierarchical learning.

- We observe there naturally exist two perspectives of hierarchies in terms of cross-modal correspondence, both on the language side and the video side.

- Motivated by the observations, we model the three-level bottom-up linguistic encoding process and the three-level coarse-to-fine temporal decoding by a hierarchical encoder and a hierarchical decoder, respectively.



Architecture



Training Loss

- Training objectives include an encoder loss, a decoder loss and a grounding loss as given below:

$$\mathcal{L}_{total} = \mathcal{L}_{enc} + \mathcal{L}_{dec} + \mathcal{L}_{grd}$$

- The encoder loss consists of three terms as formulated below:

$$\mathcal{L}_{enc} = \mathcal{L}_{enc}^{\mathcal{VW}} + \mathcal{L}_{enc}^{\mathcal{VS}} + \mathcal{L}_{enc}^{\mathcal{VP}}$$

- $\mathcal{L}_{enc}^{\mathcal{VW}}$: word-level encoder loss for **video-word** alignment
- $\mathcal{L}_{enc}^{\mathcal{VW}}$: sentence-level encoder loss for **video-sentence** alignment
- $\mathcal{L}_{enc}^{\mathcal{VW}}$: paragraph-level encoder loss for **video-paragraph** alignment

- The decoder loss consists of four terms as formulated below:

$$\mathcal{L}_{enc} = \mathcal{L}_{dec}^{\mathcal{P}} + \mathcal{L}_{dec}^{\mathcal{S}} + \mathcal{L}_{union}^{\mathcal{W}} + \mathcal{L}_{subset}^{\mathcal{W}}$$

- $\mathcal{L}_{dec}^{\mathcal{P}}$: decoding loss to regress the **paragraph-level** timestamps by ground-truth labels
- $\mathcal{L}_{dec}^{\mathcal{S}}$: decoding loss to regress the **sentence-level** timestamps by ground-truth labels
- $\mathcal{L}_{union}^{\mathcal{W}}$: weakly-supervised decoding loss to regress the union of word-level timestamps with sentence timestamps
- $\mathcal{L}_{subset}^{\mathcal{W}}$: weakly-supervised decoding loss to encourage word-level timestamps be subsets of sentence timestamps

- $\mathcal{L}_{grd}$: regress the **final grounding predictions** based on multi-level semantics with the sentence timestamps

Comparison with State-of-the-arts

- The proposed HSCNet outperforms all the existing state-of-the-art methods by introducing the hierarchical modeling framework into video paragraph grounding.
- Note that HSCNet exhibits stronger advantages when it comes to multi-modal inputs with more complicated semantic structures, e.g., longer videos and more sentences.

Method	ActivityNet-Captions				TACoS			
	R@IoU=0.3	R@IoU=0.5	R@IoU=0.7	mIoU	R@IoU=0.1	R@IoU=0.3	R@IoU=0.5	mIoU
DRN [34]	-	45.45	24.36	-	-	-	23.17	-
2D-TAN [39]	59.45	44.51	26.54	-	47.59	37.29	-	-
BPNet [30]	58.98	42.07	24.69	42.11	-	25.96	20.96	19.53
CBLN [19]	66.34	48.12	27.60	-	49.16	38.98	27.65	-
MMN [29]	65.05	48.59	29.26	-	51.39	39.24	26.17	-
SLP [18]	-	52.89	32.04	-	-	42.73	32.58	-
Beam Search [7]	62.53	46.43	27.12	-	48.46	38.14	25.72	-
3D-TPN [39]	67.56	51.49	30.92	-	55.05	40.31	26.54	-
DepNet [2]	72.81	55.91	33.46	-	56.10	41.34	27.16	-
PRVG [26]	78.27	61.15	37.83	55.62	61.64	45.40	26.37	29.18
SVPTR [12]	78.07	61.70	38.36	55.91	67.91	47.89	28.22	31.42
HSCNet	**81.89**	**66.57**	**44.03**	**59.71**	**76.28**	**59.74**	**42.00**	**40.61**

Ablation Studies

- Converting the multi-level model into a single sentence-level model leads to a significant performance degradation.
- Each modeling level is important for accurate grounding and complements well to each other.

Word Level	Sentence Level	Paragraph Level	Recall @0.3	Recall @0.5	Mean IoU
✗	✓	✗	49.27	30.14	32.66
✗	✓	✓	53.73	32.67	35.18
✓	✓	✗	54.79	38.76	37.43
✓	✓	✓	**59.74**	**42.00**	**40.61**

Ablation Studies

- The word-level cross-modal alignment has the greatest impact on the video paragraph grounding, since it injects the lowest-level details into the model.
- Aligning visual and textual representations at multiple levels during encoding is an effective way to learn fine-grained cross-modal correspondence.

Word Alignment	Sentence Alignment	Paragraph Alignment	Recall @0.3	Recall @0.5	Mean IoU
✗	✗	✗	29.98	12.81	22.38
✗	✗	✓	36.34	17.14	25.56
✗	✓	✗	48.94	31.18	32.71
✓	✗	✗	52.23	34.25	35.02
✗	✓	✓	51.66	32.86	34.46
✓	✗	✓	55.66	35.85	37.49
✓	✓	✗	56.80	38.82	38.39
✓	✓	✓	**59.74**	**42.00**	**40.61**

Ablation Studies

- Decoding the temporal timestamps at multiple semantic levels facilitates to supervise the localization learning, which contributes to better grounding performance.

Word Decoder	Sentence Decoder	Paragraph Decoder	Recall @0.3	Recall @0.5	Mean IoU
✗	✓	✗	53.56	38.55	37.57
✗	✓	✓	55.41	38.90	38.33
✓	✓	✗	56.78	40.21	39.10
✓	✓	✓	**59.74**	**42.00**	**40.61**

Qualitative Analysis

- The model is able to give reasonable results for a long untrimmed video with frames of similar visual appearances and subtle action variations.
- In terms of paragraph-level and sentence-level understanding, the model gives an accurate paragraph timestamp and predicts a series of sentence timestamps reasonably ordered within the time span of the paragraph.
- In terms of the word-level understanding, we observe many success cases. For instance, the verbs “peels” and “chops” are grounded by two moments close to the action durations and ordered in a correct relation.

Conclusion

- We propose a novel HSCNet, which is designed as a hierarchical encoder-decoder architecture to learn and exploit the multi-level visual-textual correspondence.

- We show that the combination of the encoder and decoder enables the system to learn fine-grained hierarchical cross-modal understanding capabilities for grounding.

- The proposed HSCNet establishes new state-of-the-arts on two challenging benchmarks.