電機資訊學院 2020 作 BRAIN PLUS HAND 實作專題競賽

See-Through-Text Grouping for Referring Image Segmentation

EECS02 賈松昊

Motivation

■ Design Principle

- the model includes a ConvRNN and proper weighting/attention schemes to ensure the iterative fusion process alternately improves the two aspects and boosts the final segmentation.

■ See-through-Text Grouping

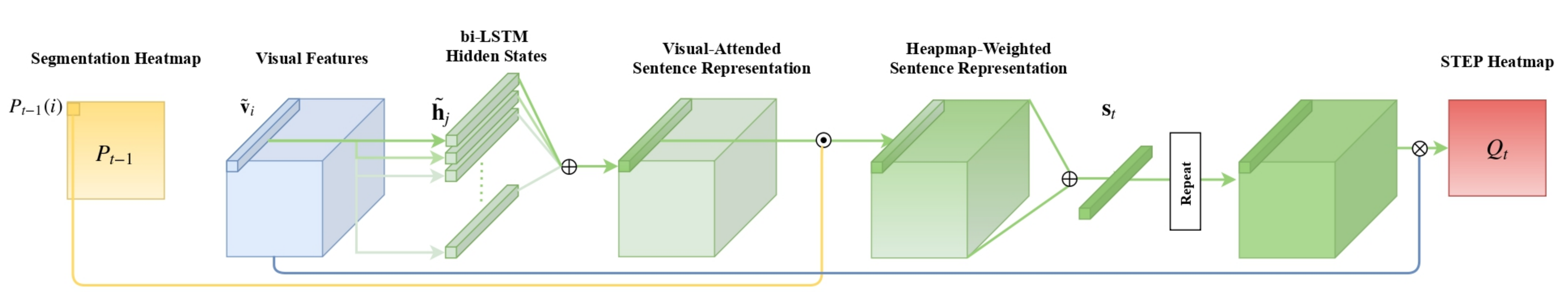
- the bottom-up grouping of pixelwise visual-textual co-embedding yields a See-through-Text Embedding Pixelwise (STEP) heatmap.
- the top-down process by ConvRNN converts the input STEP heatmap into a refined one, which is used to update the textual representation of the referring expression for the ensuing time step of ConvRNN.

Key Idea

■ The more precise the textual representation is, the better the result of referring segmentation is.

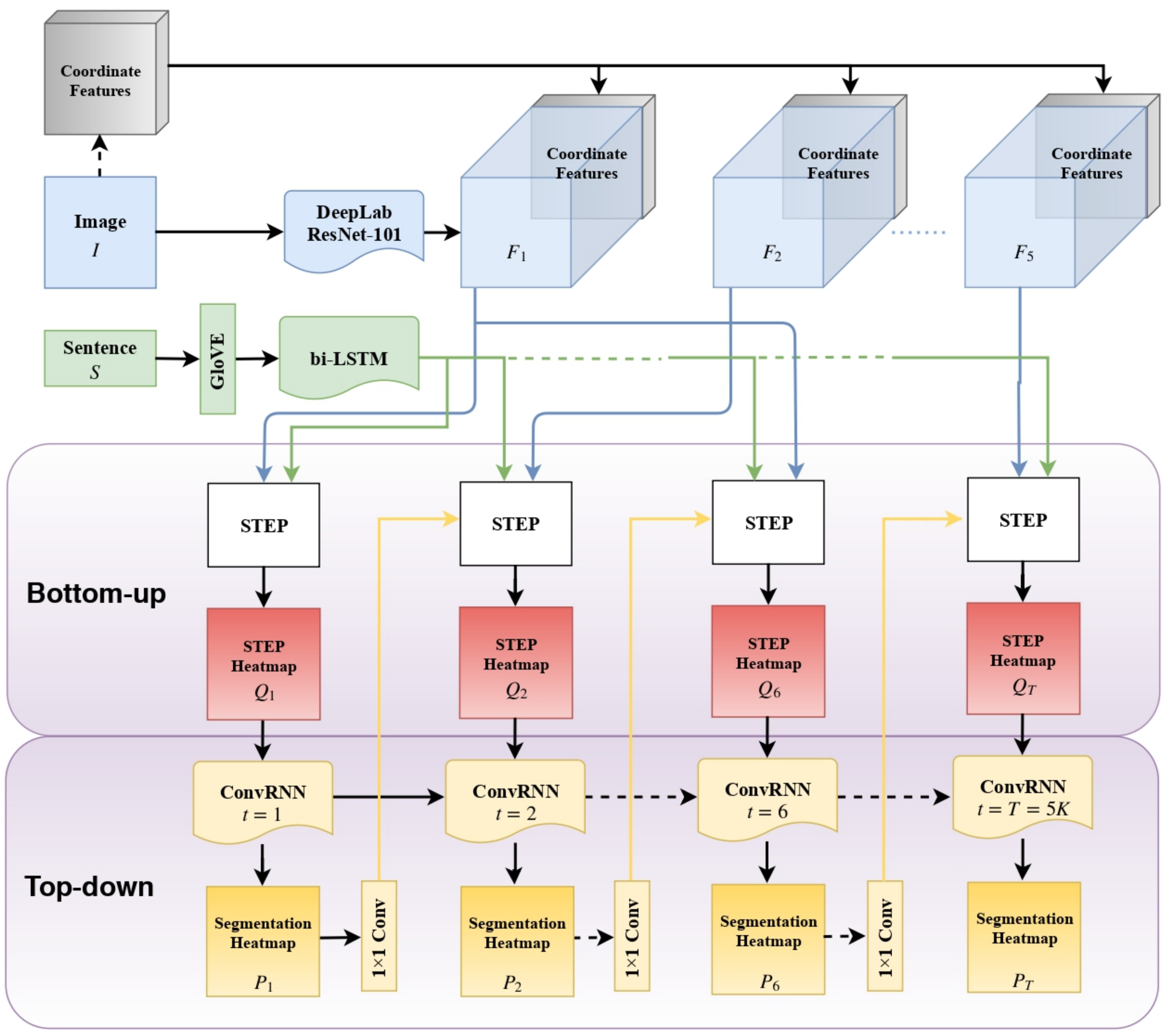
■ Reweighted Textual Representation

- our method first computes per-pixel textual representations and then yields their weighted combination by referencing the segmentation hint form the previous segmentation heatmap.



Model

■ **Our referring image segmentation model, with K-fold see-through-text grouping. The green lines indicate the language flow. The blue lines indicate the visual flow. The five sets of feature maps (from five conv layers) implies ConvRNN has 5×K time steps.**



Results

■ Experimental results of mIoU metric on four datasets.

``-'' indicates no available results.
``n/a'' denotes the method does not use the same split as the RIS ones.

Method	ReferIt	UNC			UNC+			GRef
	test	val	testA	testB	val	testA	testB	val
LSTM-CNN [1]	48.03	-	-	-	-	-	-	28.14
LSTM-CNN+ [2]	49.91	-	-	-	-	-	-	34.06
RMI+DCRF [3]	58.73	45.18	45.69	45.57	29.86	30.48	29.50	34.52
RRN+DCRF [4]	63.63	55.33	57.26	53.95	39.75	42.15	36.11	36.45
DMN [5]	52.81	49.78	54.83	45.13	38.88	44.22	32.29	36.76
KWAN [6]	59.19	-	-	-	-	-	-	36.92
MAttNet [7]	-	56.51	62.37	51.70	46.67	**52.39**	40.08	n/a
Ours (5-fold)	**64.13**	**60.04**	**63.46**	**57.97**	**48.19**	52.33	**40.41**	**46.40**

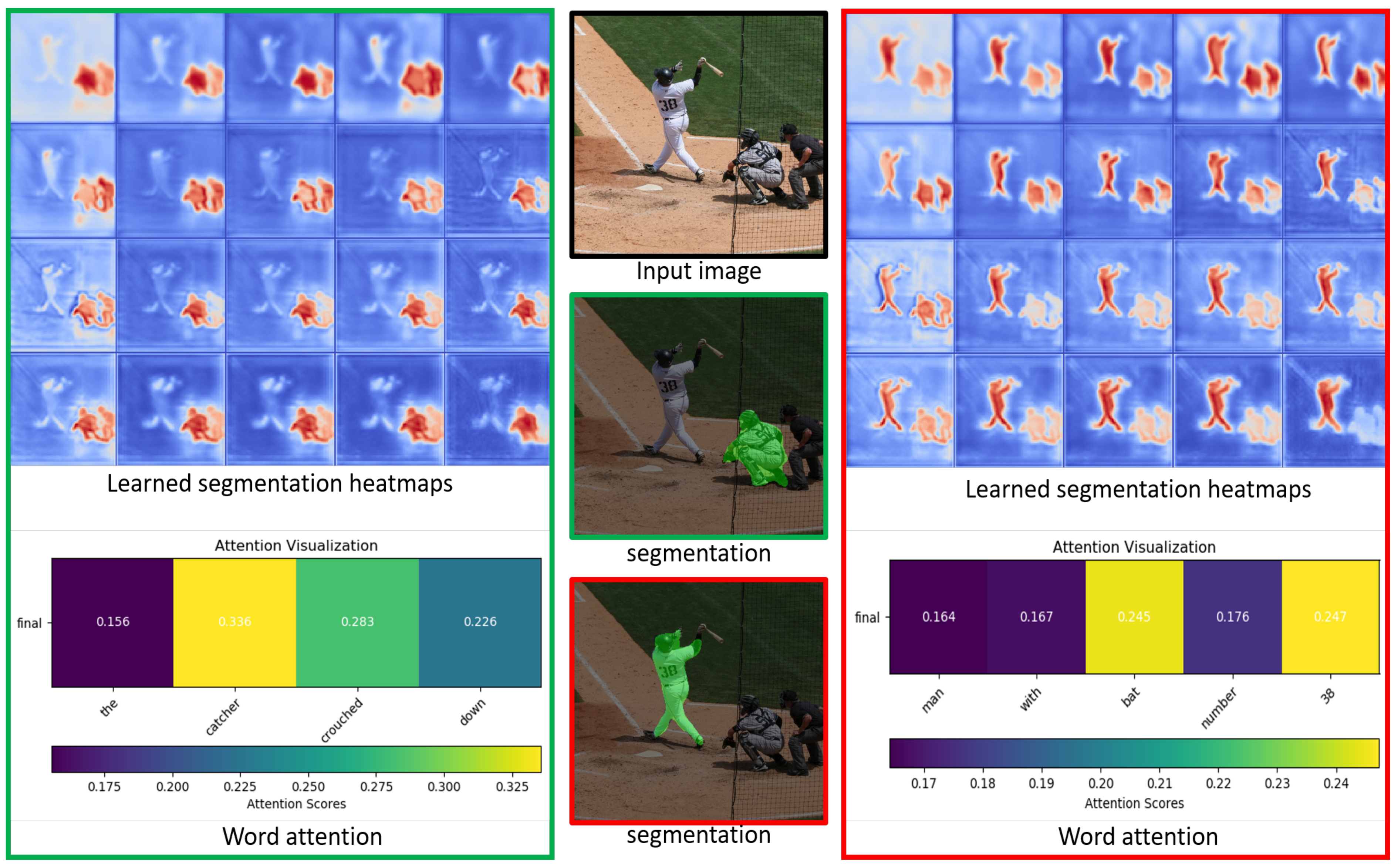
■ Ablation study

Comparison between different fold numbers, based on the Prec@X, on the UNC val split.

Version	Prec@0.5	Prec@0.6	Prec@0.7	Prec@0.8	Prec@0.9	mIoU
Ours (1-fold)	66.03	55.91	44.35	27.65	7.43	56.68
Ours (2-fold)	67.72	58.70	47.03	30.86	8.13	57.23
Ours (4-fold)	70.15	63.37	53.15	36.53	10.45	59.13

■ Qualitative results

Learned segmentation heatmaps, word attention, and the final segmentation.



Learned segmentation heatmaps

Word attention

Input image

segmentation

segmentation

Learned segmentation heatmaps

Word attention

■ References

[1] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In ECCV 2016.
[2] Ronghang Hu, Marcus Rohrbach, Subhashini Venugopalan, and Trevor Darrell. Utilizing large scale vision and text datasets for image segmentation from referring expressions. In ArXiv, 2016.
[3] Chenxi Liu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan L. Yuille. Recurrent multimodal interaction for referring image segmentation. In ICCV, 2017.
[4] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In CVPR, 2018.
[5] Edgar Margffoy-Tuay, Juan C. Pérez, Emilio Botero, and Pablo Arbeláez. Dynamic multimodal instance segmentation guided by natural language queries. In ECCV, 2018.
[6] Hengcan Shi, Hongliang Li, Fanman Meng, and Qingbo Wu. Key-word-aware network for referring expression image segmentation. In ECCV, 2018.
[7] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. Mattnet: Modular attention network for referring expression comprehension. In CVPR, 2018.

ICCV 2019 Seoul, Korea