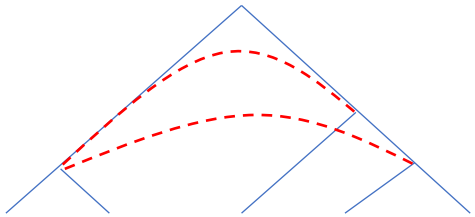
Augmenting Transformers with Recursively Composed Multi-grained Representations

Xiang Hu, Qingyang Zhu, Kewei Tu, Wei Wu

Motivation

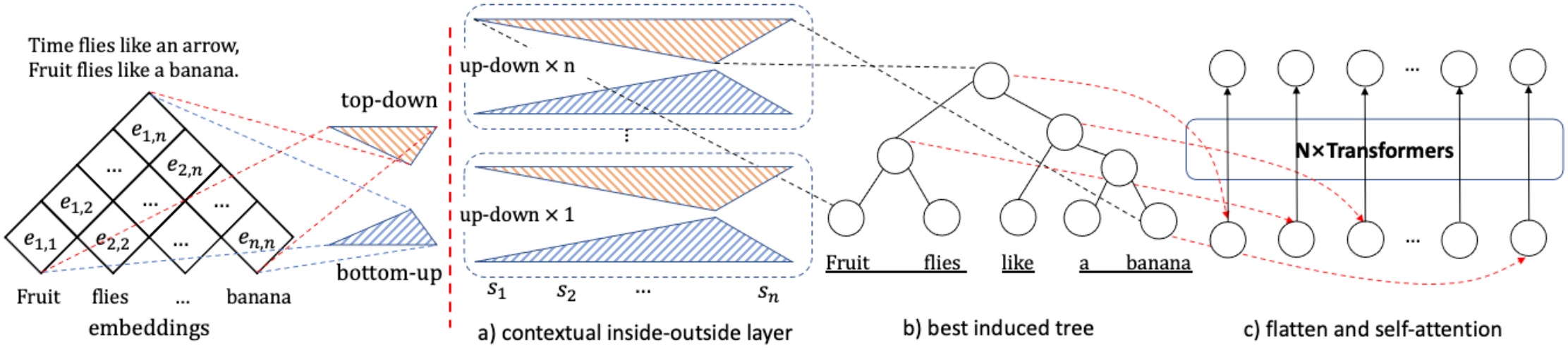
- Human language has hierarchical structures and different granularities
  - Time flies like an arrow. Fruit flies like a banana.
- Transformer perform self-attention solely at the token-level
- What if self-attention happens over multi-grained constituents? How to obtain hierarchical structures with out gold-trees?



Fruit flies like a banana

Approach

- Deep inside outside algorithm with iterative up & down



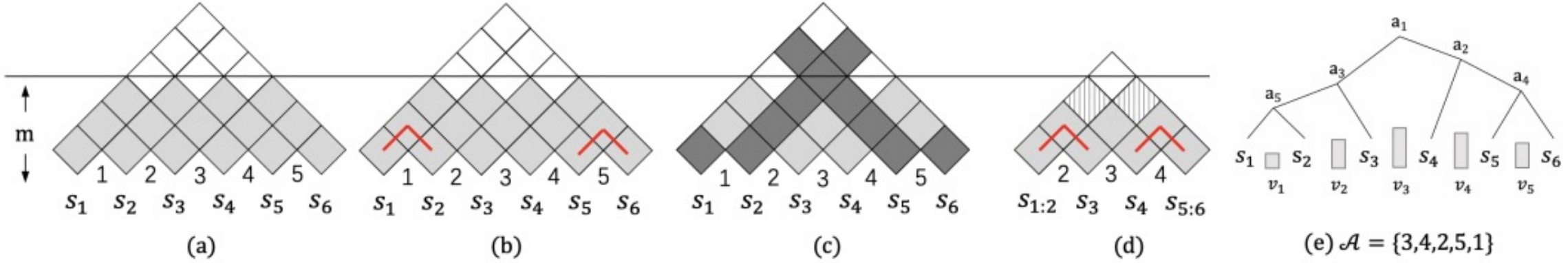
grammar induction

flatten nodes and apply self-attention

Using the masked language model as the pre–training objective

Approach

- log N variant of the deep inside-outside encoder



Experiment results(span-level tasks)

System	NEL	SRL	CTL	COREF
Fast-R2D2	91.67/92.67	79.49/80.21	91.34/90.43	87.77/86.97
Transformer6$_{\text{mean}}$	90.41/90.17	88.37/88.48	95.99/93.84	89.55/88.11
Transformer6$_{\text{max}}$	90.49/90.70	88.86/88.99	96.33/95.28	90.01/89.45
ReCAT$_{\text{share}}[3, 1, 3]$	**95.03/94.78**	**92.73/92.79**	**98.49/98.47**	**93.99/92.74**
Transformer9$_{\text{mean}}$	90.37/89.27	88.46/88.77	96.21/92.98	89.37/88.09
Transformer9$_{\text{max}}$	90.93/90.28	89.00/89.12	96.49/95.92	90.32/89.38
ReCAT$_{\text{noshare}}[3, 1, 3]$	**94.84/94.18**	**92.86/93.01**	**98.56/98.58**	**94.06/93.19**
For Reference				
BERT$_{\text{mean}}$	96.49/95.77	93.41/93.49	98.31/97.91	95.63/95.58
BERT$_{\text{max}}$	96.61/96.17	93.48/93.60	98.35/98.38	95.71/96.00

Table 2: Dev/Test performance for four span-level tasks on Ontonotes 5.0. All tasks are evaluated using F1 score. All models except BERT are pre-trained on wiki103 with the same setup.

Experiment results (GLUE)

System	natural language inference			single-sentence		sentence similarity	
	MNLI-(m/mm)	RTE	QNLI	SST-2	CoLA	MRPC(f1)	QQP
Fast-R2D2	69.64/69.57	54.51	76.49	**90.71**	**40.11**	**79.53**	85.95
Fast-R2D2+Transformer	68.25/67.30	55.96	76.93	89.10	36.06	78.09	**87.52**
DIORA*+Transformer	68.87/68.35	55.56	77.23	88.89	36.58	78.87	86.69
Parser+Transformer	67.95/67.16	54.74	76.18	88.53	18.32	77.56	86.13
Transformer×3	69.20/69.90	53.79	72.91	85.55	30.67	78.04	85.06
Transformer×6	73.93/73.99	**57.04**	79.88	86.58	36.04	80.80	86.81
$\text{ReCAT}_{\text{noshare}}[1,1,3]$	72.77/73.59	54.51	73.83	84.17	23.13	79.24	85.02
$\text{ReCAT}_{\text{share w/o iter}}$	75.03/75.32	56.32	80.96	84.94	20.86	78.86	85.36
$\text{ReCAT}_{\text{share w/o NT}}$	74.24/74.06	55.60	80.10	85.89	28.36	79.03	85.98
$\text{ReCAT}_{\text{share w/o TFM}}$	68.87/68.24	—	—	83.94	—	—	—
$\text{ReCAT}_{\text{share}}[3,1,3]$	**75.48/75.43**	56.68	**81.70**	86.70	25.11	79.45	86.10
Transformer×9	76.01/76.47	56.70	**83.20**	86.92	**36.89**	79.71	**88.18**
$\text{ReCAT}_{\text{noshare}}[3,1,3]$	75.75/75.79	**57.40**	82.01	86.80	26.69	**80.65**	85.97
$\text{ReCAT}_{\text{noshare}}[3,1,6]$	**76.33/77.12**	56.68	82.04	**88.65**	35.09	80.62	86.82
For reference							
GumbelTree†	69.50/ —	—	—	90.70	—	—	—
CRvNN†	72.24/72.65	—	—	88.36	—	—	—
Ordered Memory†	72.53/73.2	—	—	90.40	—	—	—

Table 3: Evaluation results on GLUE benchmark. The models with † are based on GloVe embeddings and their results are taken from Ray Chowdhury & Caragea (2023). The others are pre-trained on wiki103 with the same setups.

Experiment results (grammar induction)

Model	mem. cplx	PTB $F_1(\mu)$
Fast-R2D2$_{m=4}$	$O(n)$	57.22
ReCAT$_{\text{share}}[3,1,3]_{\text{m=4}}$	$O(n)$	56.07
ReCAT$_{\text{share}}[3,1,3]_{\text{m=2}}$	$O(n)$	55.11
ReCAT$_{\text{noshare}}[3,1,3]_{\text{m=4}}$	$O(n)$	**65.00**
ReCAT$_{\text{noshare}}[3,1,3]_{\text{m=2}}$	$O(n)$	64.06
ReCAT$_{\text{noshare w/o iter m=2}}$	$O(n)$	45.20
For Reference		
C-PCFG	$O(n^3)$	55.2†
NBL-PCFG	$O(n^3)$	60.4†
TN-PCFG	$O(n^3)$	64.1†
ON-LSTM	$O(n)$	47.7‡
S-DIORA	$O(n^3)$	57.6†
StructFormer	$O(n^2)$	54.0‡

Model	NNP	VP	NP	ADJP
Fast-R2D2	83.44	63.80	70.56	68.47
ReCAT$_{\text{share}}[3,1,3]_{\text{m=2}}$	77.22	67.05	66.53	69.44
ReCAT$_{\text{share}}[3,1,3]_{\text{m=4}}$	**85.41**	66.14	68.43	72.92
ReCAT$_{\text{noshare}}[3,1,3]_{\text{m=2}}$	80.36	64.38	80.93	**80.96**
ReCAT$_{\text{noshare}}[3,1,3]_{\text{m=4}}$	81.71	**70.55**	**82.94**	78.28

Table 4: Left table: F1 score of unsupervised parsing on PTB dataset. Values with † and ‡ are taken from Yang et al. (2022) and Shen et al. (2021) respectively.
Upper table: Recall of constituents.
Word-level: NNP (proper noun). Phrase-level: VP (Verb Phrase), NP (Noun Phrase), ADJP (Adjective Phrase).
Samples of deduced trees can be found in Appendix A.7.