FOVEA TRANSFORMER: Efficient Long-Context Modeling with Structured Fine-To-Coarse Attention

Ziwei He, Jian Yuan, Le Zhou, Jingwen Leng, Bo Jiang
Shanghai Jiao Tong University

Introduction

- The quadratic complexity of self-attention limits long text processing.
- Many efficient methods sparsify the attention matrix by mixing local and global attention, discarding in-between information.
- We introduce Fovea Transformer, a long-context focused transformer that efficiently capturing global dependencies in a fine-to-coarse manner.
- It outperforms existing models in two out of three long-context summarization tasks and is competitive in the third.

Results

Model	R1	R2	RL
PRIMER	**46.1**	25.2	37.9
UPER	41.4	18.7	33.8
LSG-BART	46.0	24.2	37.4
Fovea Trans.	**46.1**	**25.3**	**38.1**

Tab.1 Results for WCEP-10

Model	R1	R2	RL
BigBird	46.32	20.65	42.33
Longformer	47.00	20.20	42.90
GoSum	49.83	23.56	45.10
LongT5-xl	50.23	**24.76**	**46.67**
BART-LS	50.30	24.30	46.30
Fovea Trans. (ours)	**50.41**	24.65	46.08

Tab.2 Results for PubMed

Model	R1	R2	RL
BART-Long-Graph	49.24	18.99	23.97
LongT5-xl	48.20	19.40	24.90
PRIMER	49.90	21.10	25.90
SPADE	-	19.63	23.70
Fovea Trans. (ours)	**50.32**	**21.50**	**26.62**

Tab.3 Results for Multi-News

Methodology

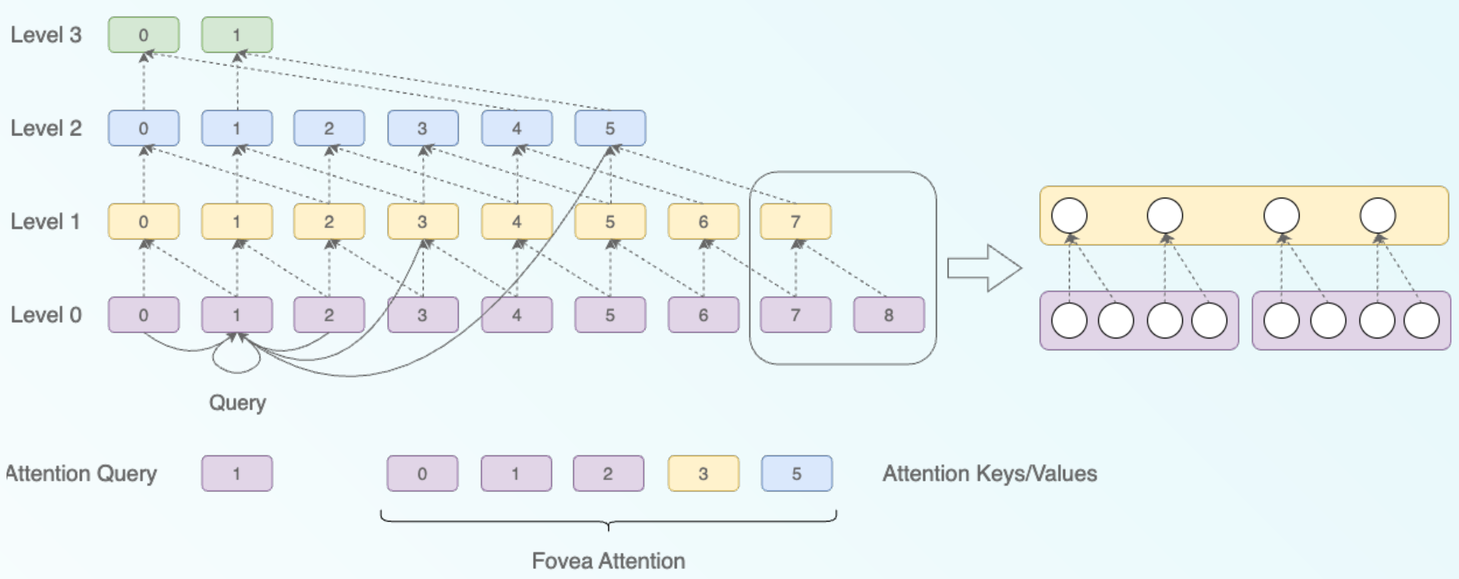

Fig.1 Illustration of tree construction and fovea attention.

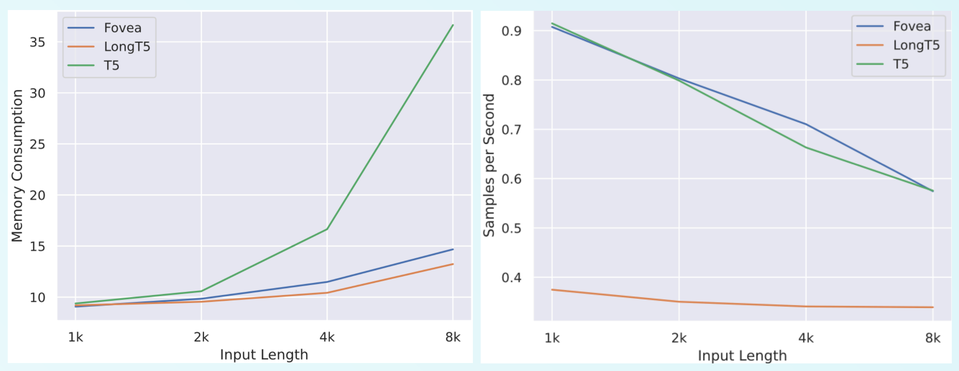

Fig.2 LongT5 and Fovea Transformer have a much smaller memory footprint compared to the regular transformer T5. However, Fovea Transformer trains significantly faster than LongT5.

Fovea Transformer, utilizing Fovea attention for multi-resolution context focus.

1. First it constructs a tree of token blocks from the input through a bottom-up process, grouping them for coarser views at higher levels (See Fig.1). For the i-th node at level q,

$$u_{q,i} = \frac{1}{2^q} \sum_{k=i}^{i+2^q-1} e_k$$

2. Fovea attention uses this tree to form self-attention components, facilitating fine-to-coarse context transitions. Formally, for the i-th query token, on its right side, the fovea attention selects the following k nodes in level q to form its kv components.

$$\{u_{q,\, i+k(2^q-1)+1+j\times 2^q} \mid j \in [0, k-1]\}$$