Sequence Encoding

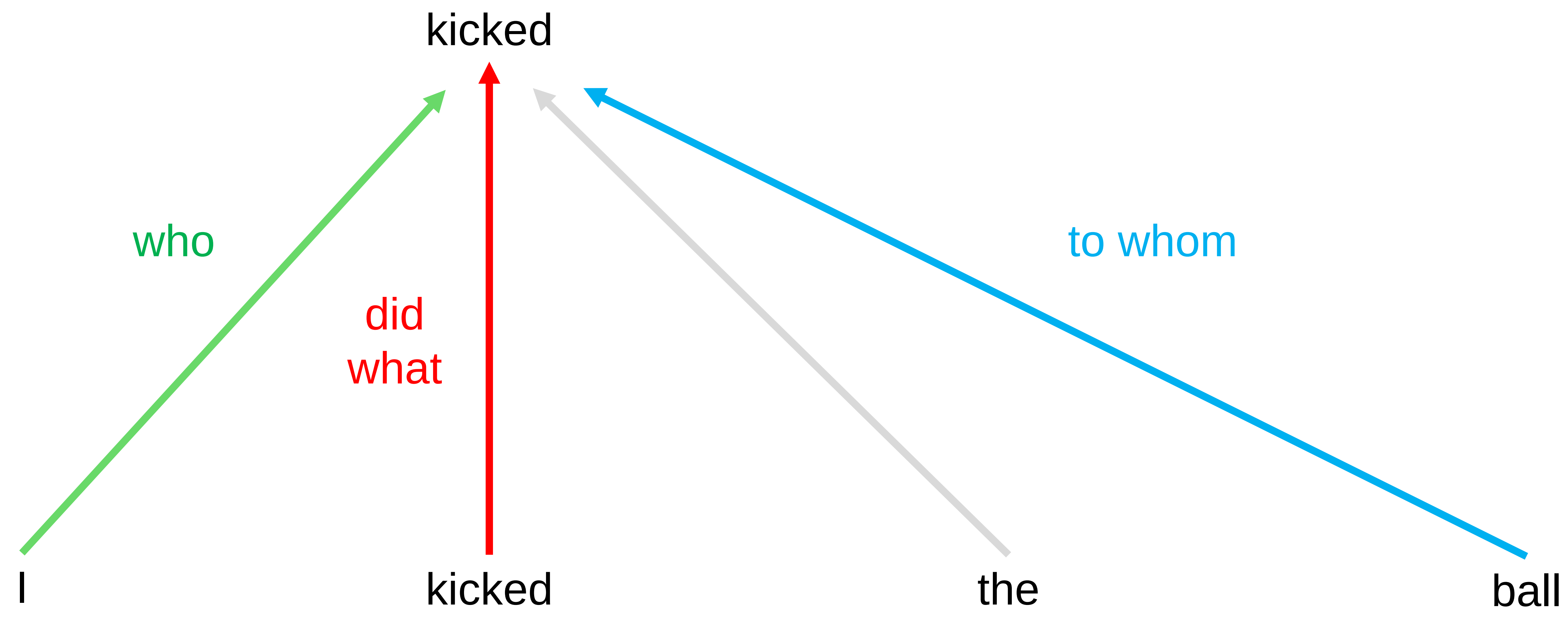
Multi-Head Attention

國立臺灣大學 資訊工程學系

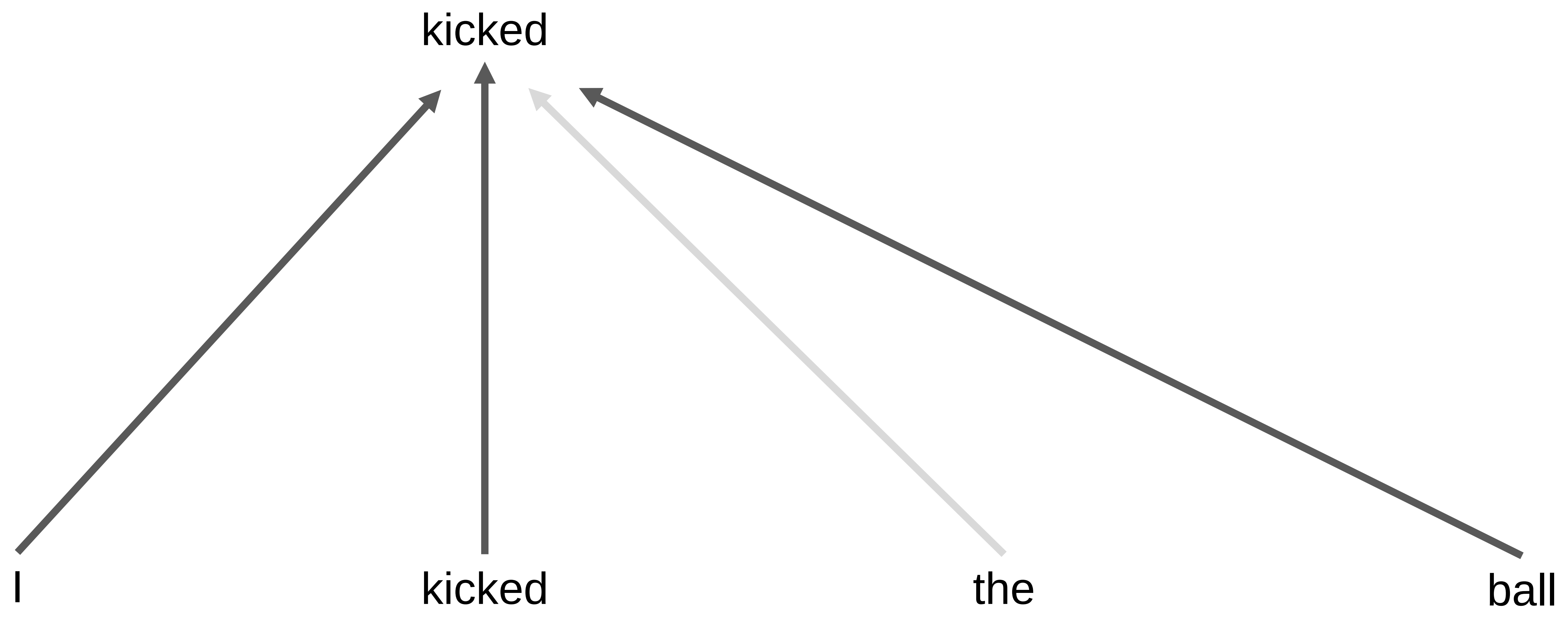
陳縕儂 助理教授

http://vivianchen.idv.tw

Convolutions



Self-Attention


kicked

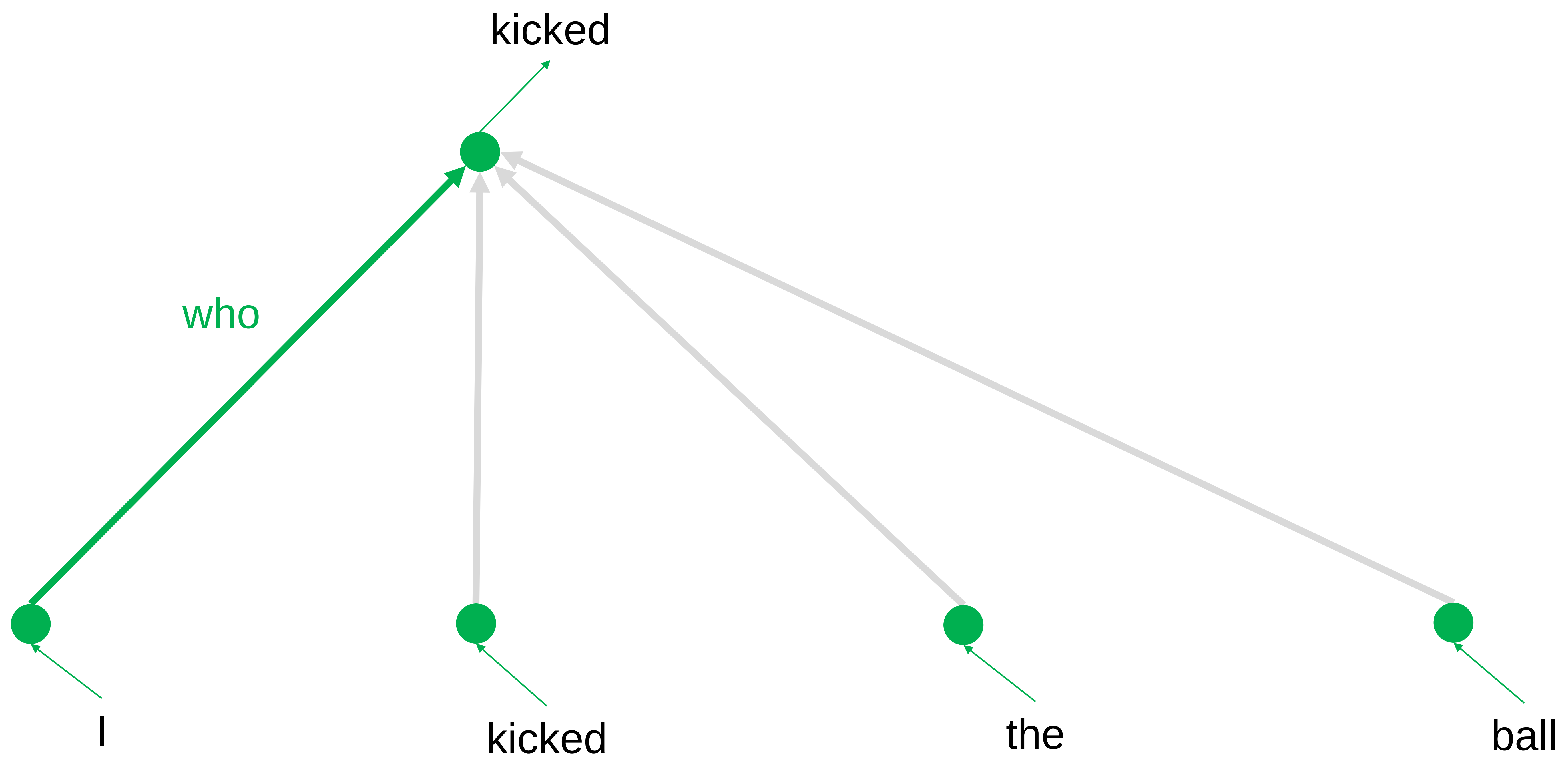
I

kicked

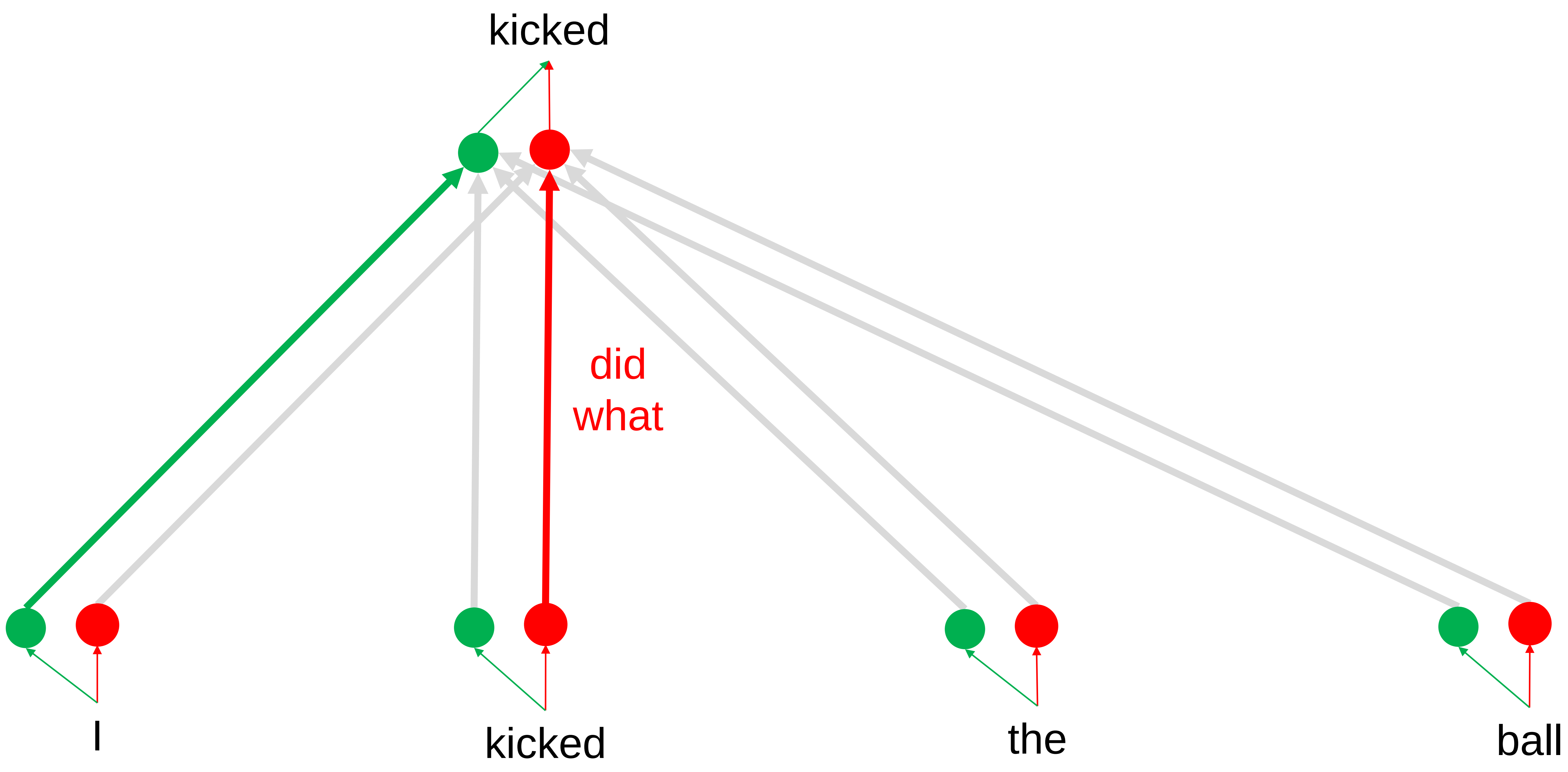
the

ball

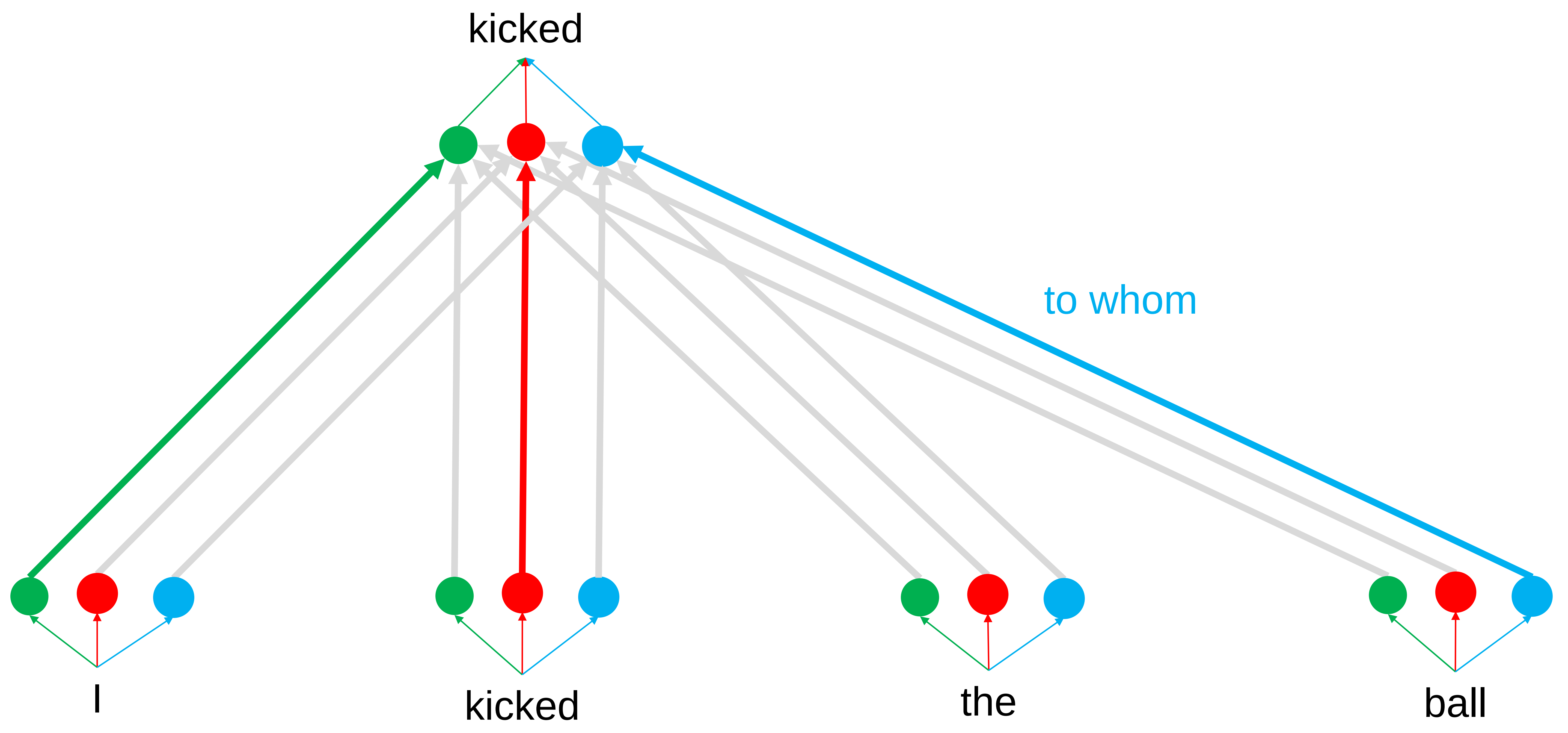
Attention Head: who



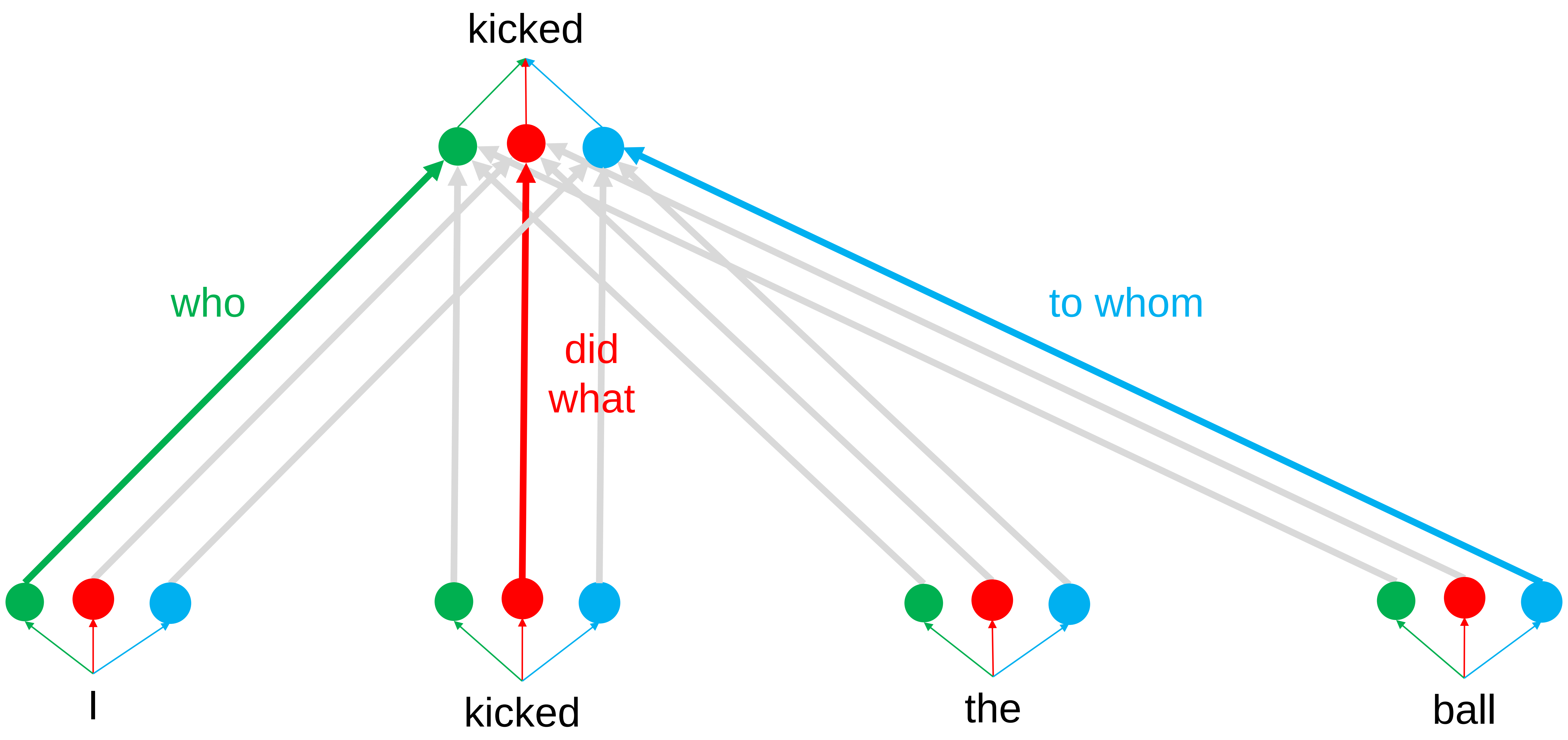
Attention Head: did what



Attention Head: to whom

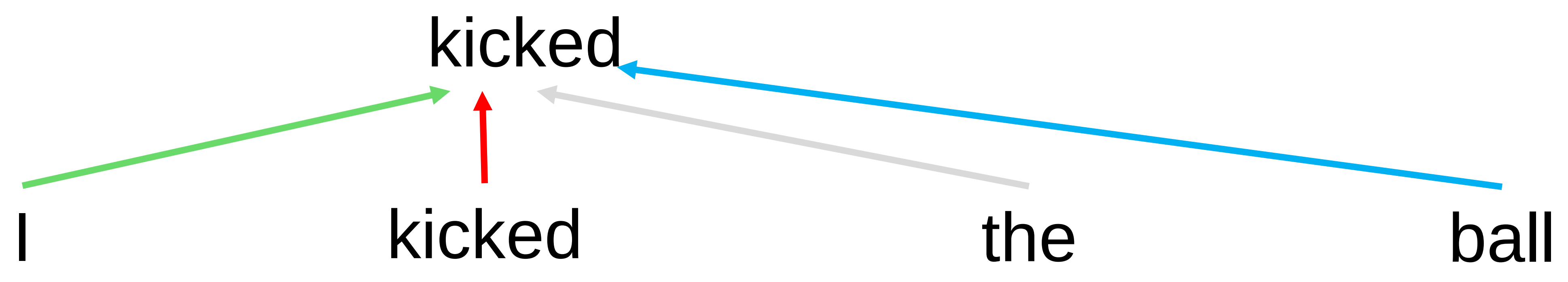


Multi-Head Attention

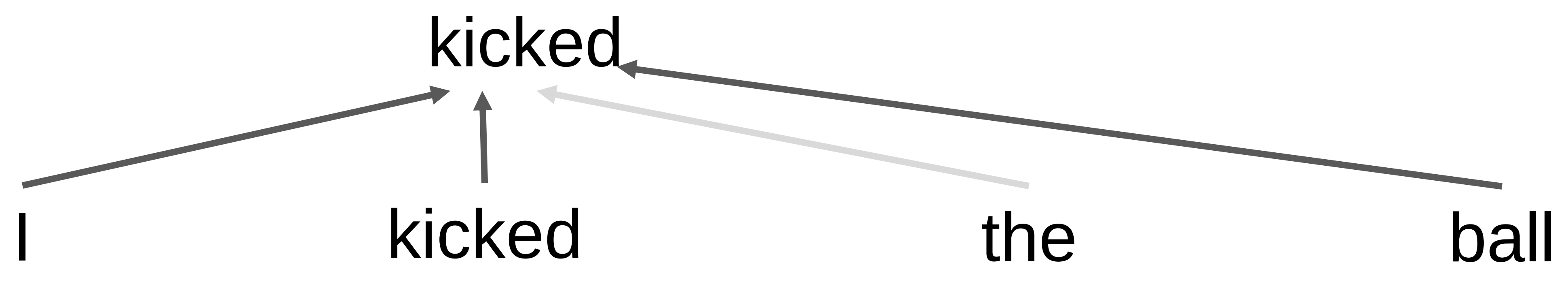


Comparison

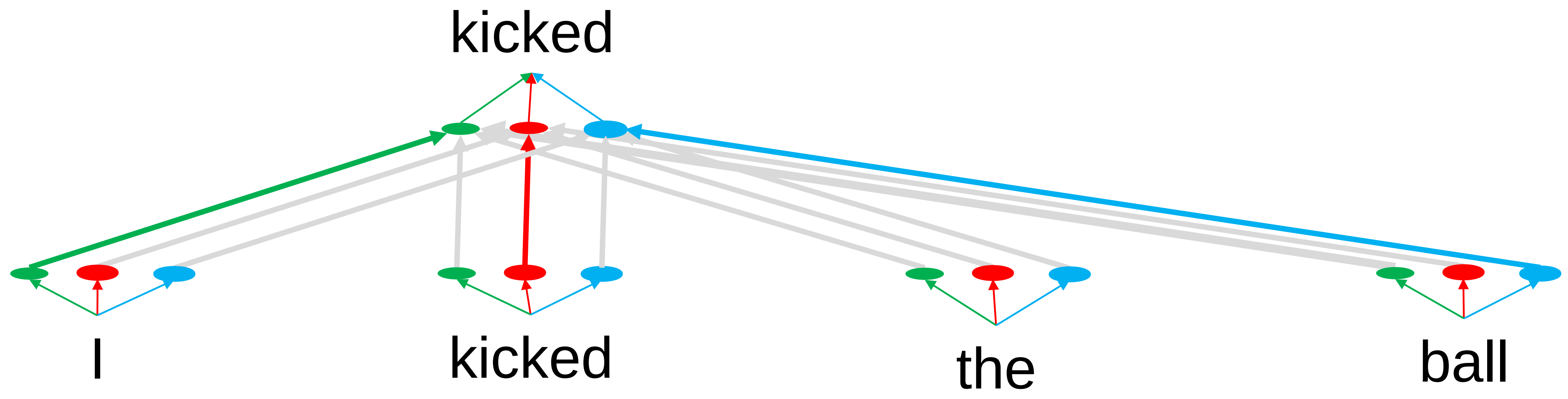
- Convolution: different linear transformations by relative positions



- Attention: a weighted average



- Multi-Head Attention: parallel attention layers with different linear transformations on input/output


kicked
I
kicked
the
ball