## Paper

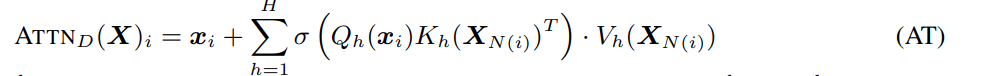
<https://browse.arxiv.org/pdf/2205.14135.pdf>

## Idea

* Combines both Random attention, window attention, and global attention

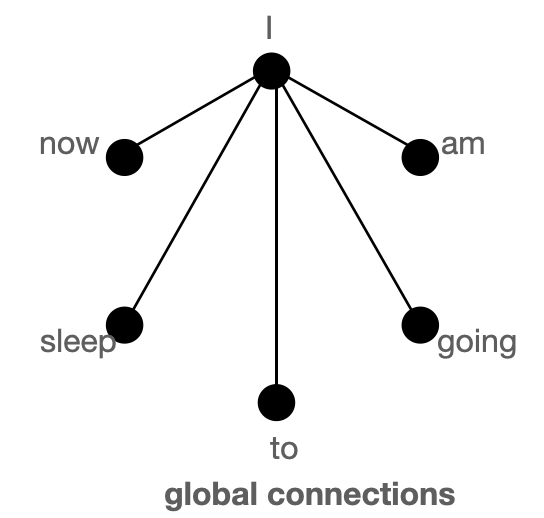
## Architecture

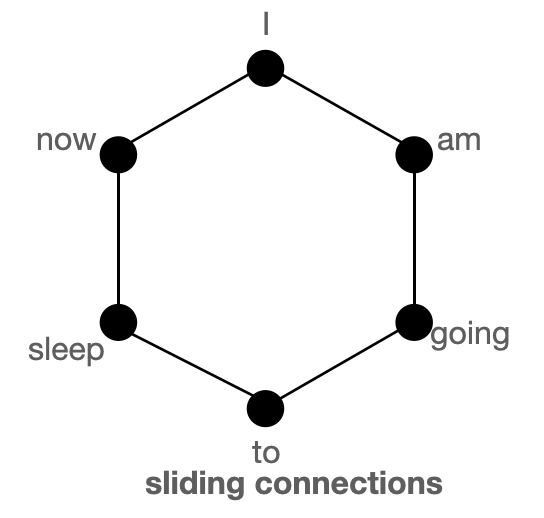
* On input sequence X = (x1 … x\_n)
* And graph D whose vertex set is [n] = {1… n}

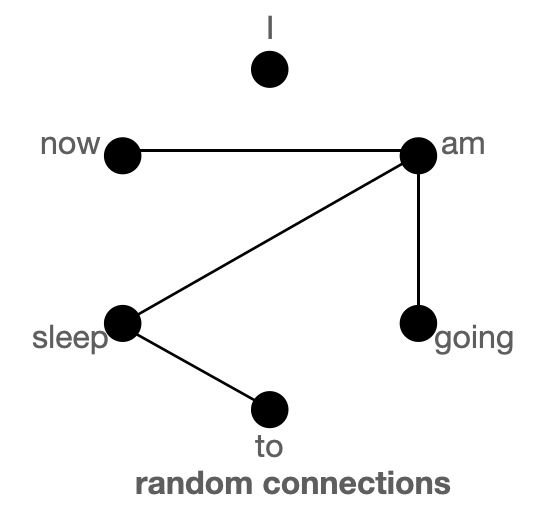


* This is just the standard attention equation
* Explains through graphs

## Understanding the need for global, sliding, random keys with Graphs







## Where is it applied

* Encoder side only

## Proof

* Dhfksd

## Sources

* <https://www.youtube.com/watch?v=WVPE62Gk3EM&ab_channel=YannicKilcher>
  + Shows attention
  + Explains star graph explanation
* <https://huggingface.co/blog/big-bird>
  + Explains graph

## Reading Group Oct 8, 2023

* Why using the graphs?
  + Its easier to explain using graphs vs normal matrices ]
  + “the quadratic complexity of self-attention can now be seen as a graph sparsification problem”
  + Erdos Renyi method
  + Can represent a graph through a Random walk of the graph?
  + Hugging face global atten graph diagram is wrong
* Is bigBird attention speed lower than original attention?
  + Not rly, its o(n)
* How is Big bird accuracy (knowing its a universal approximator) given equal sequence length?
  + It appears the sequence length varies in their experiments (as in model A has smaller seq length and they test it against BigBird with bigger context)
* Turing Completeness
  + As in is Big Bird a finite state machine and turing complete ?
  + 