Sub-quadratic Time Algorithms?

- Unfortunately seems to be no!
    - Exact or entrywise approximation
- Alman and Song show output of attention is hard to approximate to high precision under reasonable complexity assumptions

Generalizations of Softmax Attention