Self-Attention Motivation (2/2)

- SANs (Self-Attentive Networks) can access any position in constant time.

	Operations	Sequential Steps	Memory
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n \cdot d)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(n \cdot d)$
Self-attentive	$O(n^2 \cdot d)$	$O(1)$	$O(n^2 \cdot d)$

- Sequence length n, state dimensionality d, kernel size k.
- Assuming infinitely many GPU cores (or rather ALU), operations can be run in parallel, but may depend on each other, needing some Sequential Steps.