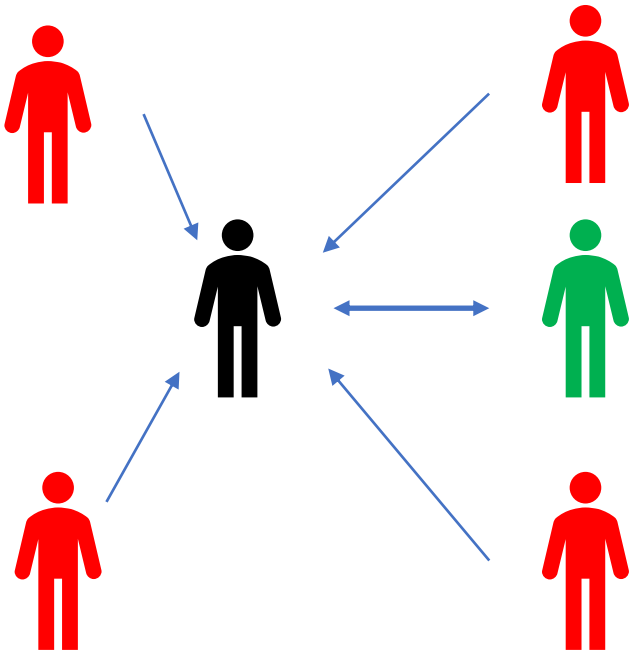
Attention

Attention

Problem

- need to focus on one aspect while ignoring other (distracting) aspects at the same time
- only keep necessary information
- ignore information rather than increase information



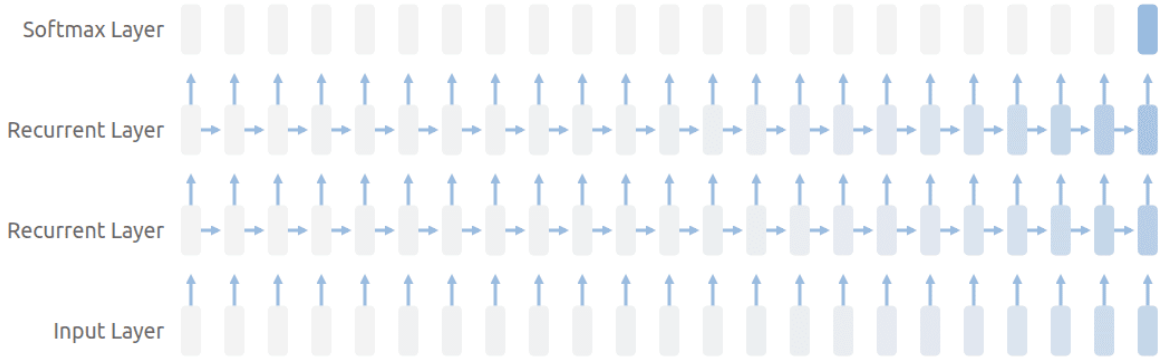
Cocktail Party Problem

Attention

Problem

- Vanishing gradient problem in RNNs
- caused by time dependency
- theoretically, every part of previous sequence know,
- BUT gradients become smaller
- farther away content has less impact
- RNNs work only well for short sequences

This is a very long sample text in which the start is forgotten.



Vanishing Gradient: where the contribution from the earlier steps becomes insignificant in the gradient for the vanilla RNN unit.

Source: https://distill.pub/2019/memorization-in-rnns/

Attention

Intuition

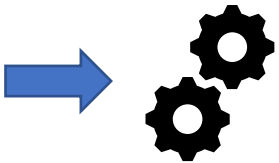
- ability to focus on certain parts of input
- Jacobian matrix represents sensitivity of output to input
- partial derivatives (back propagation)

$$J = \begin{bmatrix} \frac{\delta y_1}{\delta x_1} & \cdots & \frac{\delta y_1}{\delta x_k} \\ \cdots & \cdots & \cdots \\ \frac{\delta y_m}{\delta x_1} & \cdots & \frac{\delta y_k}{\delta x_n} \end{bmatrix}$$

Class Activation Map
(Source: https://glassboxmedicine.com/2019/06/11/cnn-heat-maps-class-activation-mapping-cam/)

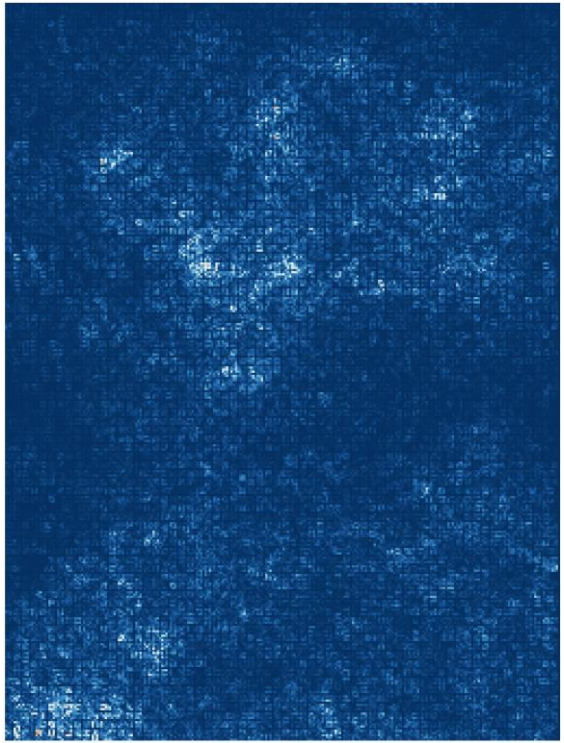
Attention

Intuition



CNN

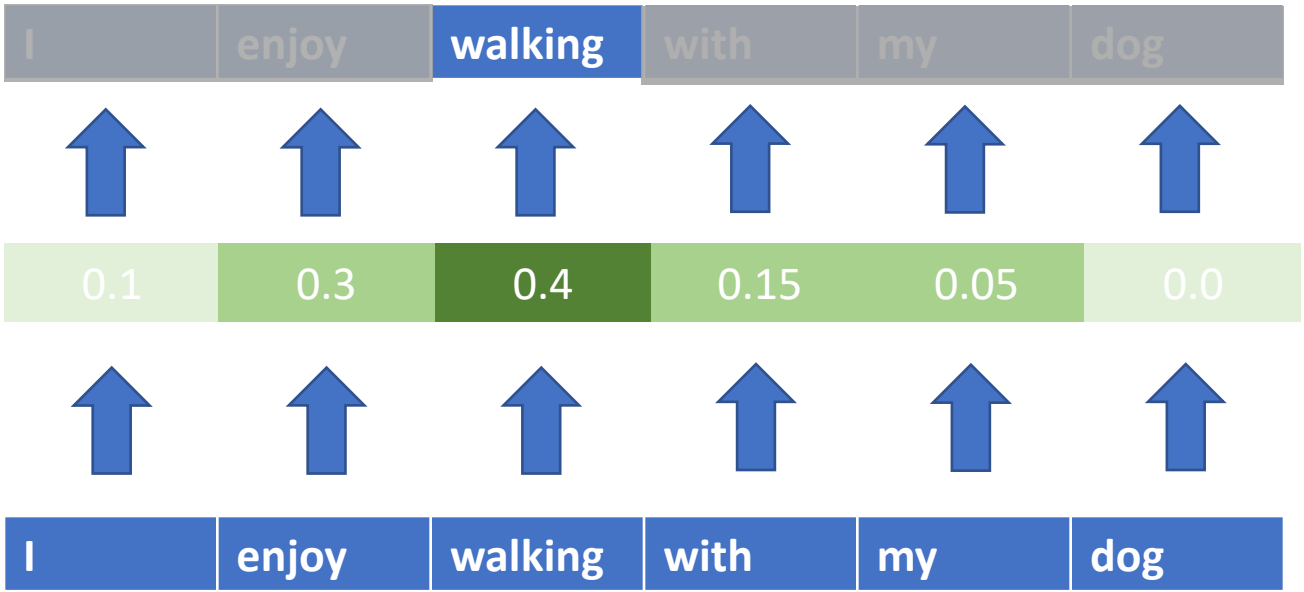
Jacobian
Matrix



Attention

Introduction

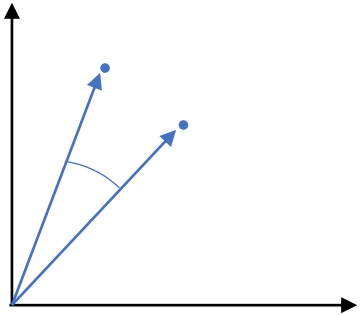
- Attention



Attention

Types of Attention

- Dot-product attention
  - calculates how close two vectors align in terms of point directions

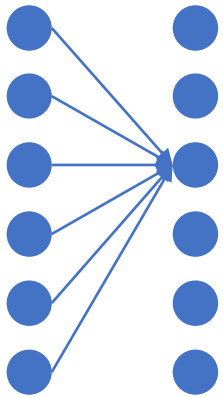


- scaled dot-product attention
  - scales dot-product by square root of key dimension
- multi-head attention
  - splits query, key, and value vectors into multiple heads and applies dot-product independently
- self-attention
  - input sequence applied as query and key

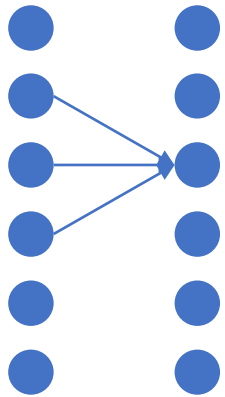
Attention

Global vs. Local Attention

- global attention – references to all input nodes
- local attention – references to subset of input nodes



Global Attention

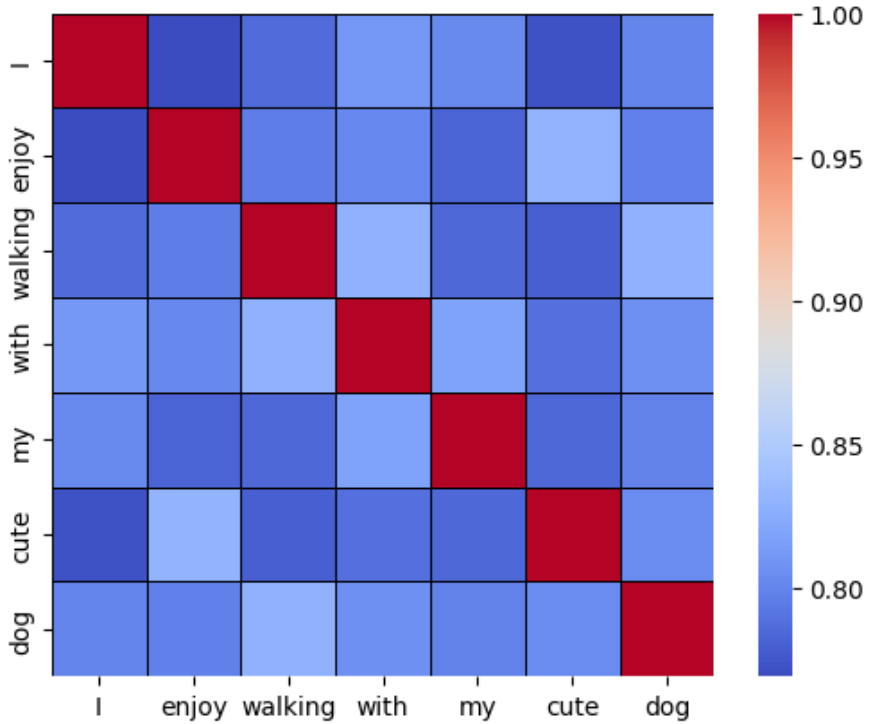


Local Attention

Attention

Self-Attention

- score between elements of the same sequence



Attention

Advantages / Disadvantages

- work a bit like skip-connections
- solves vanishing gradient problem
- improved accuracy
- improved efficiency (reduced training time)
- improved explainability

- increased training difficulty
- large amount of data required
- prone to overfitting

Attention

Applications

- Natural Language Processing
- Computer Vision, e.g.Vision Transformers
- Speech recognition (cocktail party problem)
- music generation