

Generative Pretrained Transformer

Evolution and its Future

Jaideep Ganguly, Sc.D. (MIT)

July 12, 2023

Attention Is All You Need

Ashish Vaswani*

Google Brain

avaswani@google.com

Noam Shazeer*

Google Brain

noam@google.com

Niki Parmar*

Google Research

nikip@google.com

Jakob Uszkoreit*

Google Research

usz@google.com

Llion Jones*

Google Research

llion@google.com

Aidan N. Gomez* †

University of Toronto

aidan@cs.toronto.edu

Lukasz Kaiser*

Google Brain

lukaszkaiser@google.com

Illia Polosukhin* ‡

illia.polosukhin@gmail.com

175 Billion parameters and counting!

Attention

He went to the bank and learned of his empty account, after which he went to a river bank and cried.

- ① Attention mechanism has an infinite reference window
- ② In contrast, RNN has a short reference window, LSTM has a longer window.

Word Embedding

dog →



$$\begin{bmatrix} 0.37 \\ 0.99 \\ 0.01 \\ 0.08 \end{bmatrix}$$

AJ's **dog** is a cutie

AJ looks like a **dog**

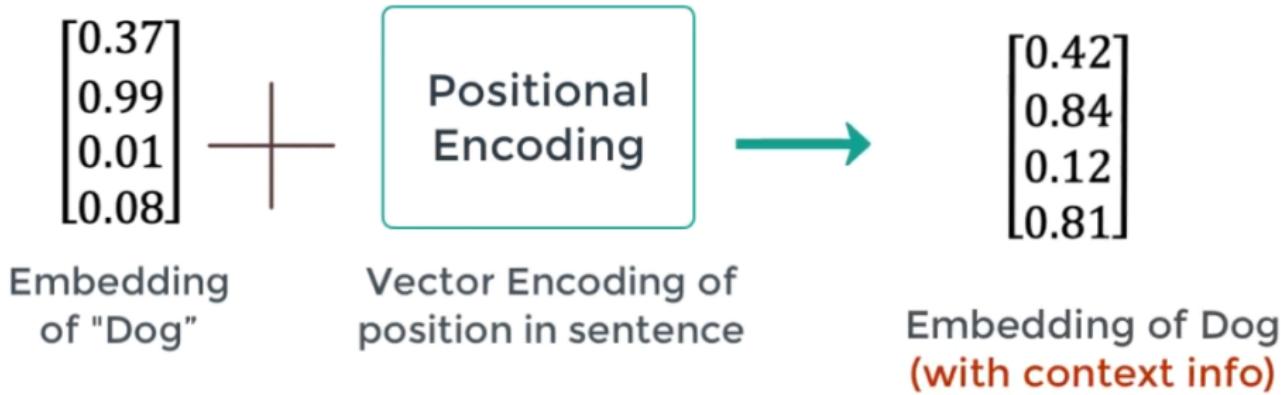
Positional Encoding (odd pos sin, even pos cos)

AJ's **dog** is a cutie → Position 2

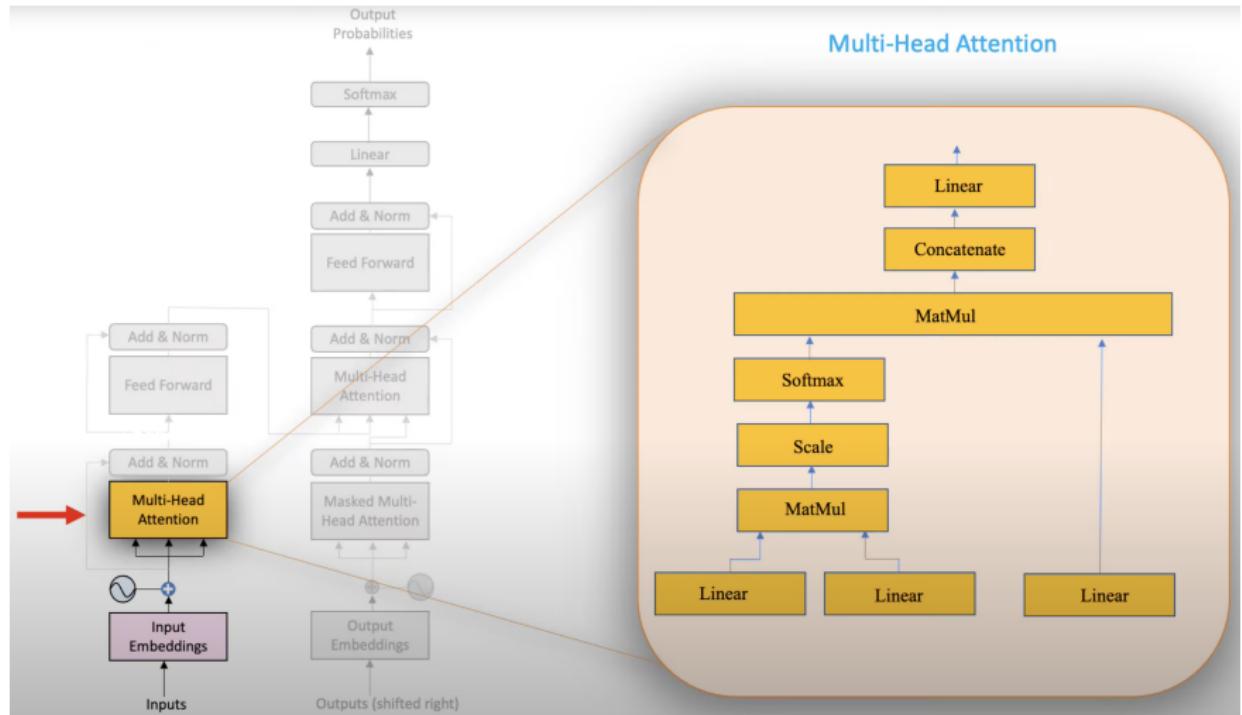
AJ looks like a **dog** → Position 5

$$PE_{(pos,2i)} = \sin(pos/10000^{2i/d_{\text{model}}})$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{2i/d_{\text{model}}})$$



Multi Head Attention



Attention: Query, Key, Value

Concept of:

- ① Query (Q)
- ② Key (K)
- ③ Value (V)

All constructed from the embedding.



I came from your other question [Self-attention original work?](#) The key/value/query formulation of attention is from the paper [Attention Is All You Need](#).

31



How should one understand the queries, keys, and values



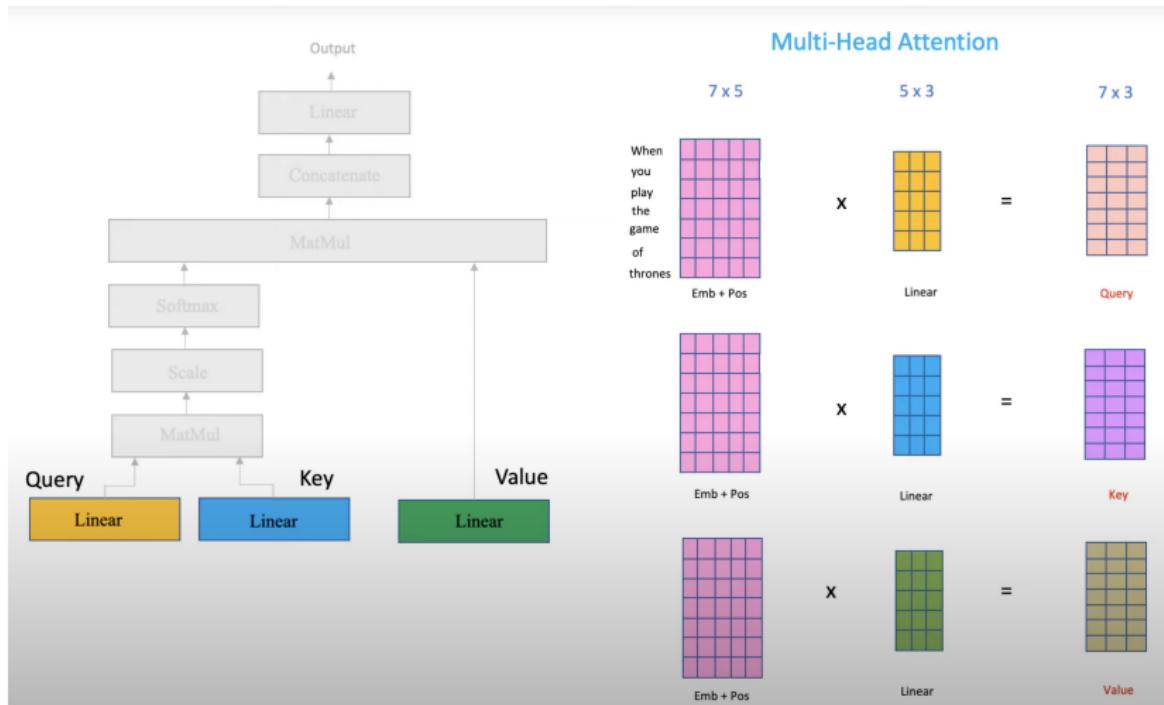
The key/value/query concepts come from retrieval systems. For example, when you type a query to search for some video on Youtube, the search engine will map your **query** against a set of **keys** (video title, description etc.) associated with candidate videos in the database, then present you the best matched videos (**values**).

+50

Mimic the retrieval of a value v_i for a query q based on a key k_i in DB.

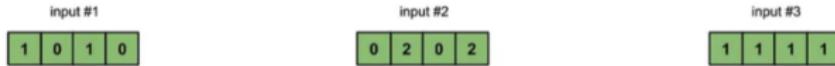
$$\text{attention}(q, k, v) = \sum_i \text{similarity}(q, k_i) \times v_i \text{ (weighted)}$$

Multi Head Attention



Attention: Q, K, V and weights Q_w , K_w , V_w

Weights are initialised randomly using a random distribution. Example:



Because every input has a dimension of 4, each set of the weights must have a shape of 4x3.
Weights are initialised randomly, it is done once before training.

Weights for key Weights for query Weights for value

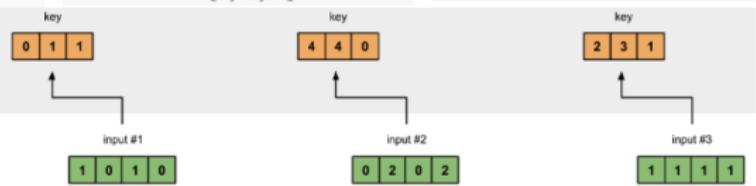
| | | |
|--|--|--|
| $\begin{bmatrix} 0, 0, 1 \\ 1, 1, 0 \\ 0, 1, 0 \\ 1, 1, 0 \end{bmatrix}$ | $\begin{bmatrix} 1, 0, 1 \\ 1, 0, 0 \\ 0, 0, 1 \\ 0, 1, 1 \end{bmatrix}$ | $\begin{bmatrix} 0, 2, 0 \\ 0, 3, 0 \\ 1, 0, 3 \\ 1, 1, 0 \end{bmatrix}$ |
|--|--|--|

Key representation
for input 1:

$$\begin{aligned} & [0, 0, 1] & [0, 0, 1] & [0, 0, 1] \\ & [1, 0, 1, 0] \times [1, 1, 0] = [0, 1, 1] & [0, 2, 0, 2] \times [1, 1, 0] = [4, 4, 0] & [1, 1, 1, 1] \times [1, 1, 0] = [2, 3, 1] \\ & [0, 1, 0] & [0, 1, 0] & [0, 1, 0] \\ & [1, 1, 0] & [1, 1, 0] & [1, 1, 0] \end{aligned}$$

Key representation:
(Vectorise)

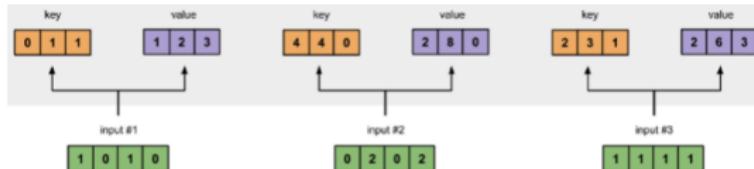
$$\begin{aligned} & [0, 0, 1] & [0, 1, 1] \\ & [1, 0, 1, 0] \times [1, 1, 0] = [0, 1, 1] & [0, 2, 0, 2] \times [0, 1, 0] = [4, 4, 0] \\ & [0, 2, 0, 2] \times [0, 1, 0] = [4, 4, 0] & [1, 1, 1, 1] \times [1, 1, 0] = [2, 3, 1] \\ & [1, 1, 1, 1] \times [1, 1, 0] = [2, 3, 1] & [1, 1, 0] \end{aligned}$$



Attention: Q, K, V and weights Q_w , K_w , V_w

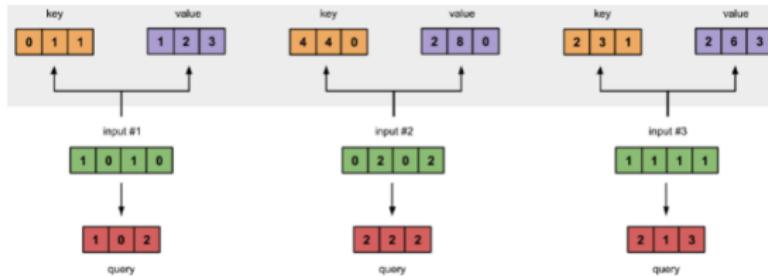
Value representation:

$$\begin{array}{ccc} [0, 2, 0] & [0, 3, 0] & [1, 2, 3] \\ [1, 0, 1, 0] \times [1, 0, 3] = [2, 8, 0] & & \\ [0, 2, 0, 2] & [1, 1, 0] & [2, 6, 3] \\ [1, 1, 1, 1] & [1, 1, 0] & \end{array}$$



Query representation:

$$\begin{array}{ccc} [1, 0, 1] & [1, 0, 0] & [1, 0, 2] \\ [1, 0, 1, 0] \times [0, 0, 1] = [2, 2, 2] & & \\ [0, 2, 0, 2] & [0, 1, 1] & [2, 1, 3] \\ [1, 1, 1, 1] & [0, 1, 1] & \end{array}$$

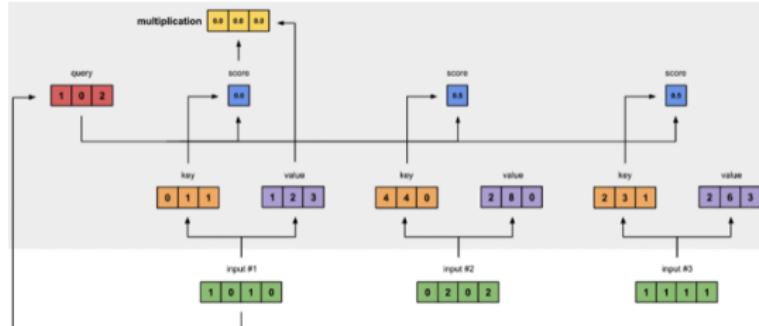


Attention: Q, K, V and weights Q_w , K_w , V_w

Attention scores - dot product between Input 1's query (red) with all keys (orange), including itself

$[0, 4, 2]$
 $[1, 0, 2] \times [1, 4, 3] = [2, 4, 4]$
 $[1, 0, 1]$

$\text{softmax}([2, 4, 4]) = [0.0, 0.5, 0.5]$



Multiply Softmax attention scores for each input (blue) by its corresponding value (purple).

1: $0.0 * [1, 2, 3] = [0.0, 0.0, 0.0]$
2: $0.5 * [2, 8, 0] = [1.0, 4.0, 0.0]$
3: $0.5 * [2, 6, 3] = [1.0, 3.0, 1.5]$

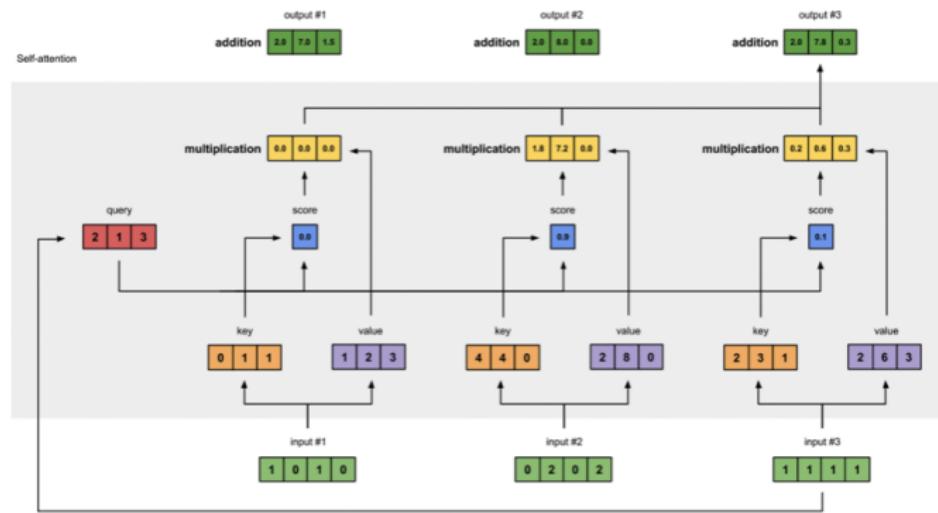
Attention: Q, K, V and weights Q_w , K_w , V_w

Weights are initialised randomly using a random distribution. Example:

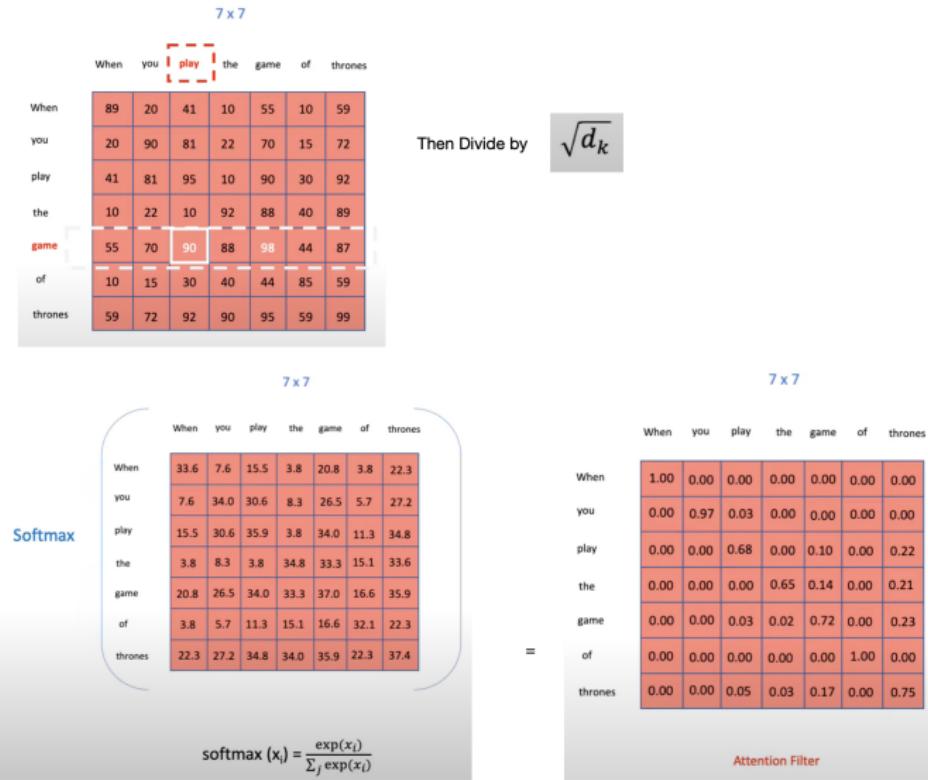
Sum weighted values to get Output 1

$$\begin{aligned} & [0.0, 0.0, 0.0] \\ + & [1.0, 4.0, 0.0] \\ + & [1.0, 3.0, 1.5] \\ \hline = & [2.0, 7.0, 1.5] \end{aligned}$$

Repeat for
Input 2 &
Input 3

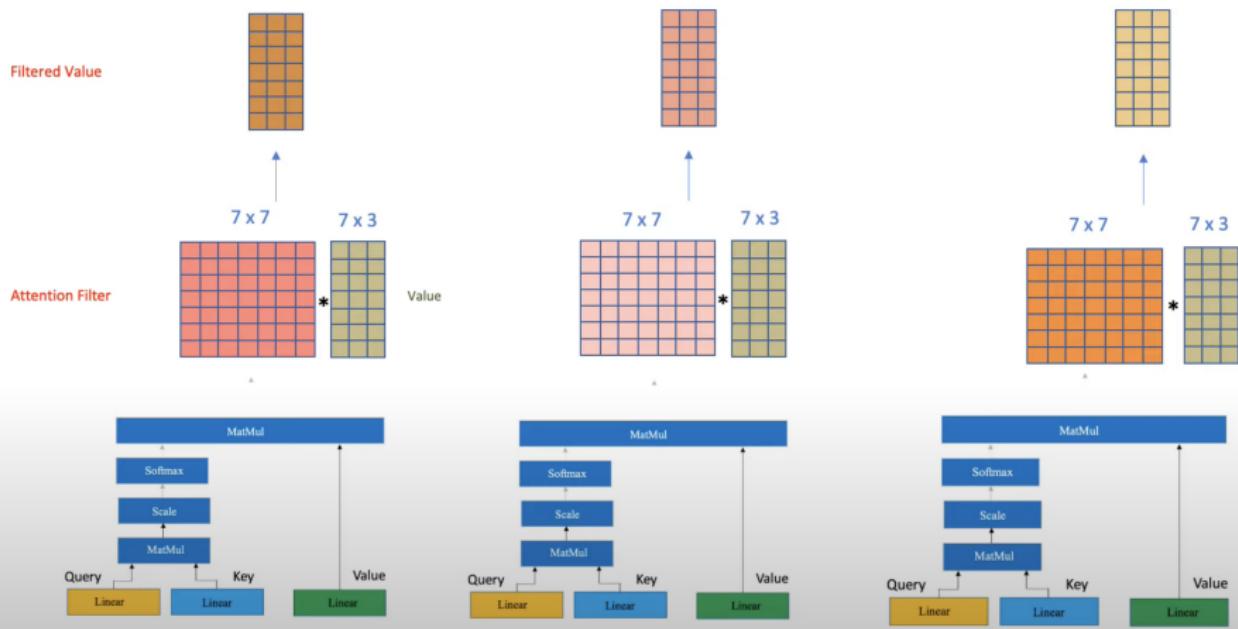


Attention Filter - filter out unnecessary information (noise)

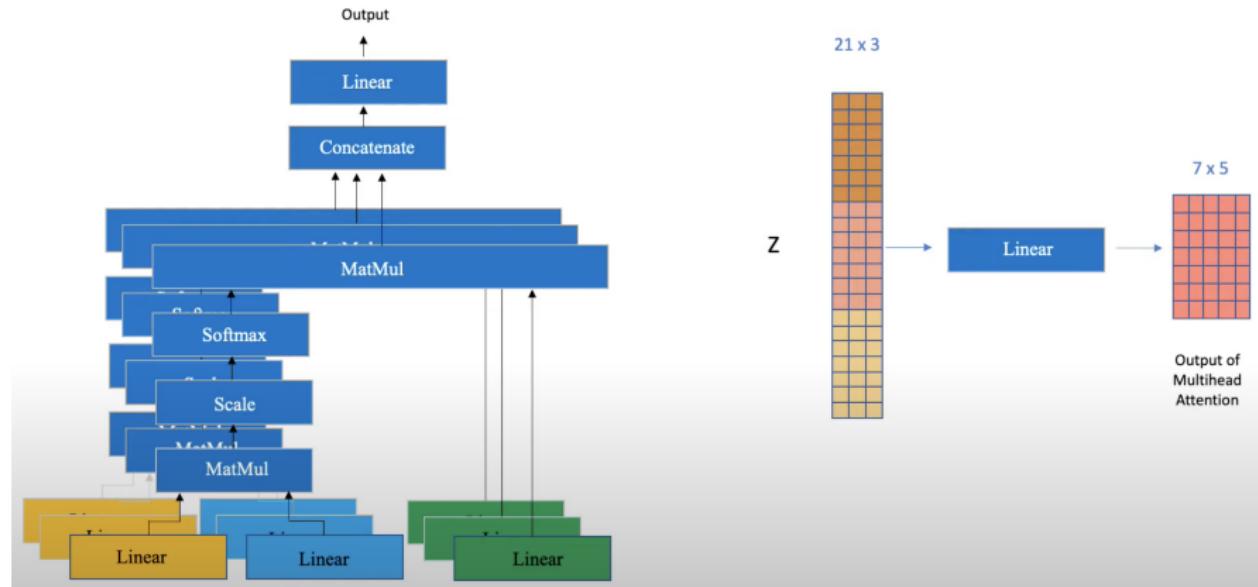


Multi Head Attention

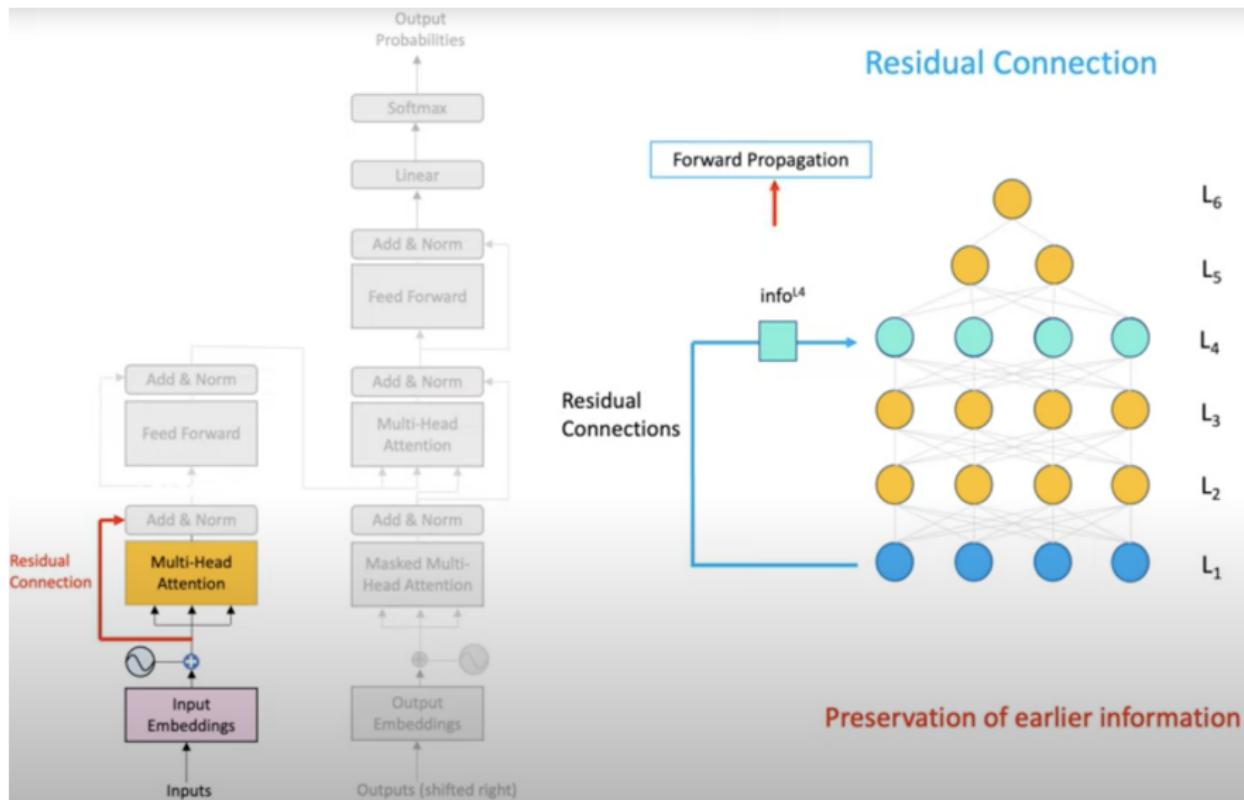
Multi-Head Attention



Multi Head Attention - Concatenation



Information Preservation



Add (Preserve Information)



Normalization

7 x 3

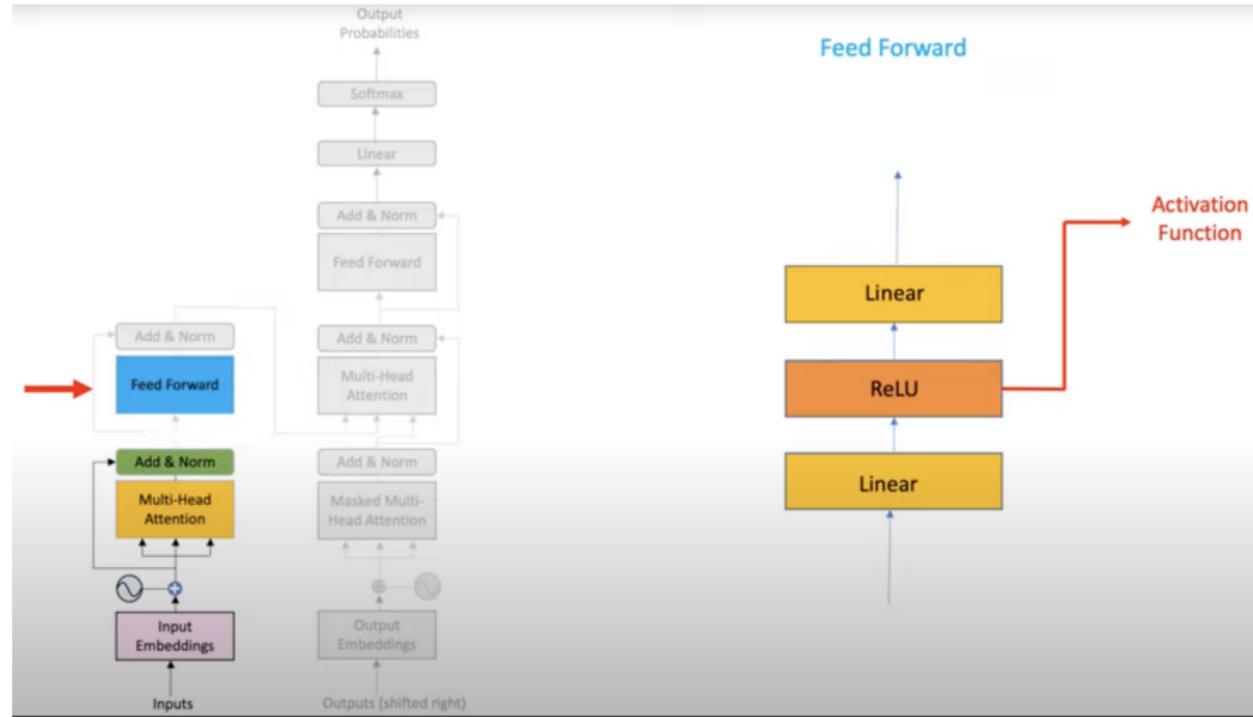
mean (μ) std (σ)

| | | | | | | | |
|------------------------|------|------|------|------|------|------|------|
| $x_0 = \text{When}$ | 0.98 | 1.28 | 0.41 | 0.27 | 0.41 | 0.67 | 0.44 |
| $x_1 = \text{you}$ | 0.52 | 0.01 | 2.06 | 0.27 | 0.33 | 0.64 | 0.82 |
| $x_2 = \text{play}$ | 2.22 | 0.27 | 0.10 | 0.41 | 2.06 | 1.01 | 1.04 |
| $x_3 = \text{the}$ | 0.99 | 1.00 | 0.11 | 0.27 | 0.33 | 0.54 | 0.42 |
| $x_4 = \text{game}$ | 0.52 | 0.01 | 0.33 | 2.06 | 0.52 | 0.69 | 0.79 |
| $x_5 = \text{of}$ | 0.10 | 2.06 | 0.73 | 0.27 | 0.41 | 0.71 | 0.79 |
| $x_6 = \text{thrones}$ | 0.33 | 0.01 | 0.13 | 0.27 | 1.28 | 0.40 | 0.51 |

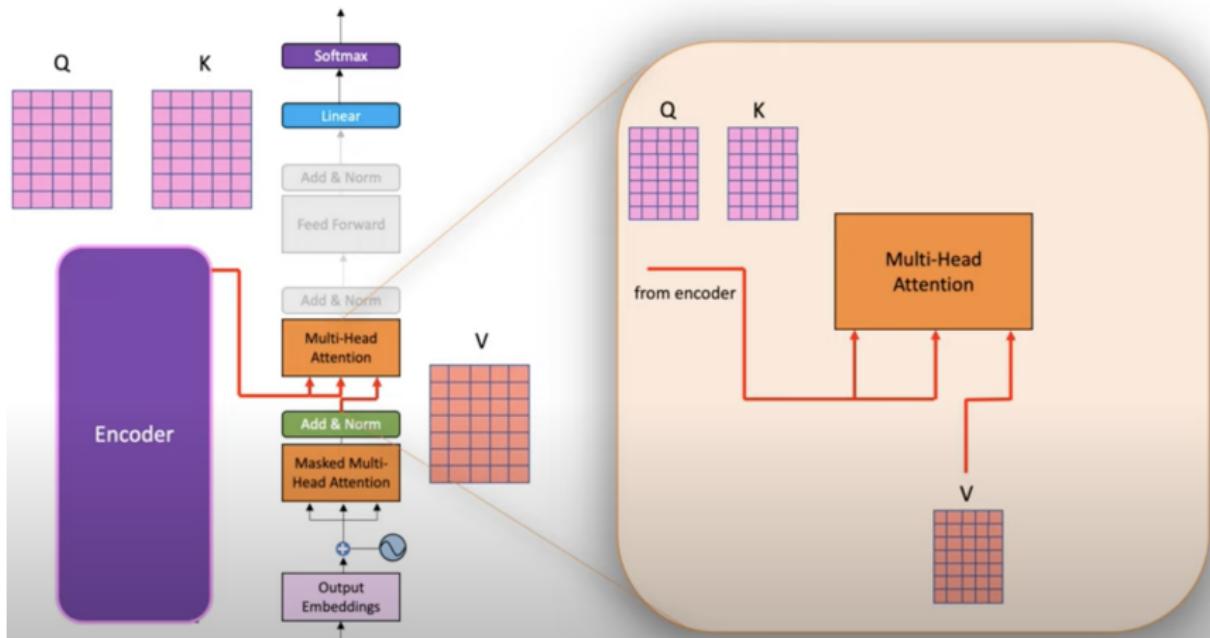
 $f_0 \quad f_1 \quad f_2 \quad f_3 \quad f_4$

$$x_0 = \frac{0.98 - 0.67}{\sqrt{0.44^2 + 0.0001}}$$

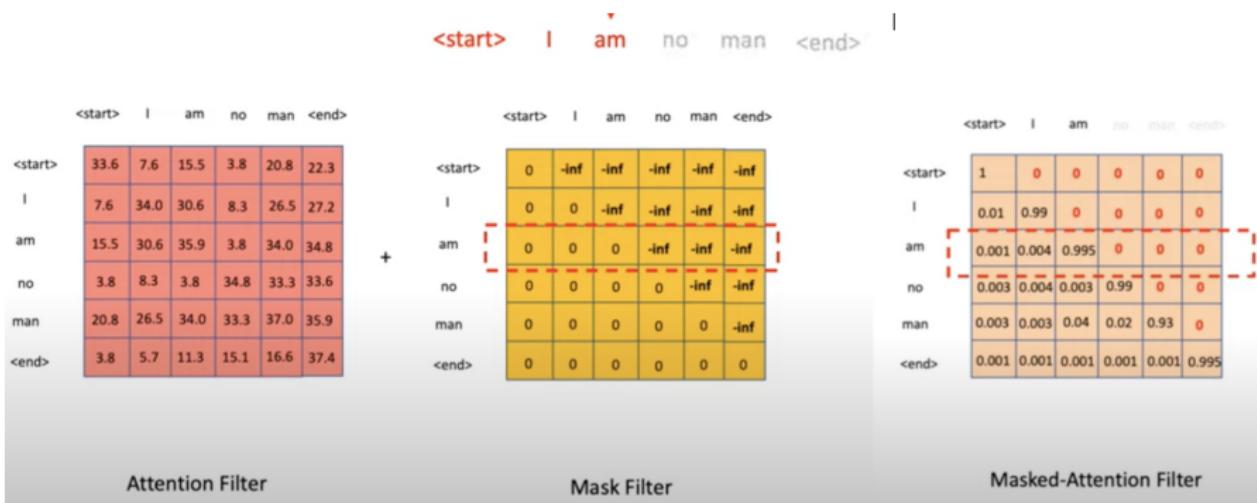
Feed Forward



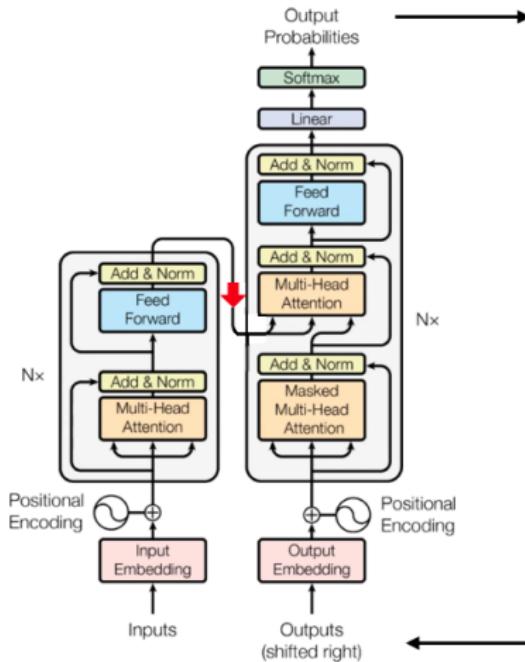
Decoder Layer



Masking



Encoder - Decoder



Decoders are autoregressive models;
They are trained to predict the next token
after reading the preceding ones

Basic Concepts

Linear Regression

- ① **Regression** is a statistical approach to find the relationship between variables between X_i and Y_i .
- ② A common function used to model training data is a **linear regression model**. The **model** or the **hypothesis** is given by:

$$y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \dots + \theta_n x_{in} \quad (1)$$

where x_{ij} is the observed value of the feature x_j , y_i is the predicted value of the outcome and θ_i are constants the values of which need to be determined. For now, we limit the the values of the features x_{ij} to numbers, positive or negative. Later on, we will study techniques on how to deal with the situation where the feature value is a string.

$$\text{Squared Error Loss (L)} = \frac{1}{2} \times \sum_{i=1}^r (\hat{y}_i - y_i)^2 \quad (2)$$

This loss function is always positive, there is a minimum, is monotonically increasing from that minimum value in both positive and negative directions. Such a function is called a **convex function**.

- ③ We need to minimize some loss function over the training data.

Gradient Descent

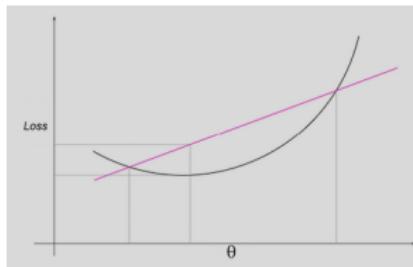


Figure: Convex Function

- ① Minimize by setting the partial derivative of the loss wrt θ_j to zero.

$$\frac{\partial L}{\partial \theta_0} = \sum_{i=1}^r (\hat{y}_i - y_i) = 0 \quad \frac{\partial L}{\partial \theta_j} = \sum_{i=1}^r (\hat{y}_i - y_i) x_{ij} = 0 \quad (3)$$

- ② Randomly assigning values to θ_j . For a convex function it does not matter what the initial weights are as it will always converge. \hat{y} and x_{ij} are observed and y_i is computed. This means the slope of the loss function can be readily computed.

Gradient Descent

Now for a given θ_j , and keeping all other θ_i where $i \neq j$ constant, we change the value of θ_j slightly and compute a new value of θ_j

$$\begin{aligned}\Delta\theta_j &= \alpha\theta_j \\ \theta'_j &= \theta_j + \Delta\theta_j\end{aligned}$$

where α is small constant and is called the **learning rate**. Since the slope of $\frac{\partial L}{\partial \theta_j}$ is already calculated from the previous equations, we can compute the new value of the loss as follows:

$$L' = L - \frac{\partial L}{\partial \theta_j}(\alpha\theta_j)$$

If $L' < L$, we continue incrementing θ_j as long as the loss decreases. When the direction changes and the loss increases, we have found the θ_j for which the loss is minimum. If $L' > L$, we travel in the reverse direction and follow the same process. Like this, we compute the value of each and every θ_j .

SoftMax

- ① In the training data the values corresponding to the features x_j can have numerical values of different magnitudes. For example, one feature may have values ranging between 0 ... 10, while another feature may have values ranging between 10,000 ... 100,000. Some of these values can also be negative and the components will obviously not add up to 1.
- ② The softmax function is a function that takes as input a vector of K real numbers, and normalizes it into a probability distribution consisting of K probabilities in the interval 0...1 with the components adding up to 1.

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^K e^{z_j}} \quad (4)$$

The feature values are normalized with the softmax function prior to inputting in a regression model.

One Hot Encoding

- ① In the linear regression and the logistic regression models that we studied so far can only deal with numerical data. It cannot handle features that have text as values. In **one-hot encoding**, the string encoded variable is replaced with new variables of boolean type. For example, a particular feature, say color, can have "red", "green" and "blue" as possible values. This feature color, will be replaced by 3 features, red, blue and green which hold the values 1 or 0. We can then encode color, in terms of red, blue and green as follows:

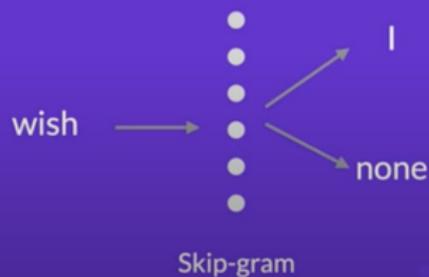
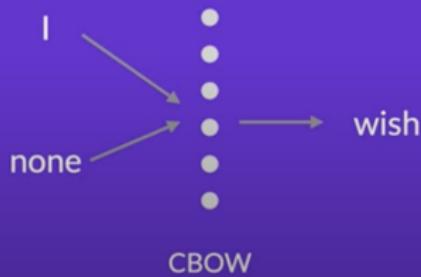
| red | green | blue |
|-----|-------|------|
| 1 | 0 | 0 |
| 0 | 1 | 0 |
| 0 | 0 | 1 |

Note that if the boolean value of the feature is 1, it must be 0 for all the other features

Word Embedding



... I wish none of this has happened. So do all who live to see such times. But that is not up to us to decide. All we have to decide is what to do with the time that is given to us...



Information & Uncertainty

- ① In Claude Shannon's information theory, one bit of information reduces the uncertainty by 2. Similarly, if 3 bits of information are sent, then the reduction in uncertainty by 2^3 , i.e., 8. This is intuitive. With 3 bits, there could be 8 possible values and so if a particular set of bits are transmitted, 8 possibilities are eliminated with 1 certainty.
- ② **Information Content** When the information is probabilistic, the self-information I_x , or Information Content of *measuring a random variable X as outcome x is defined as:*

$$I_x = \log\left(\frac{1}{p(x)}\right) = -\log(p(x)) \quad (5)$$

where $p(x)$ is probability mass function.

- ③ **Shannon Entropy** of the random variable X is defined as:

$$H(X) = \sum_x -p(x)\log(p(x)) = E(I_x) \quad (6)$$

It is the *expected information content* of the measurement of X .

Cross Entropy - Kullback–Leibler (KL) divergence

- ① Cross-Entropy is defined as:

$$H(p, q) = - \sum_i p(i) \log_2 q(i) \quad (7)$$

where p is the true distribution and q is the predicted distribution. If the predictions are perfect, then the cross-entropy is same as the entropy. If the prediction differs, then there is a divergence which is known as *Kullback–Leibler (KL) divergence*. Hence,

$$\text{Cross Entropy} = \text{Entropy} + \text{KL Divergence}$$

or,

$$\text{KL Divergence} = H(p, q) - H(p)$$

Hence, if the predicted distribution is closer to true distribution when KL divergence is low.

Neural Net, RNN, LSTM

Deep Neural Net

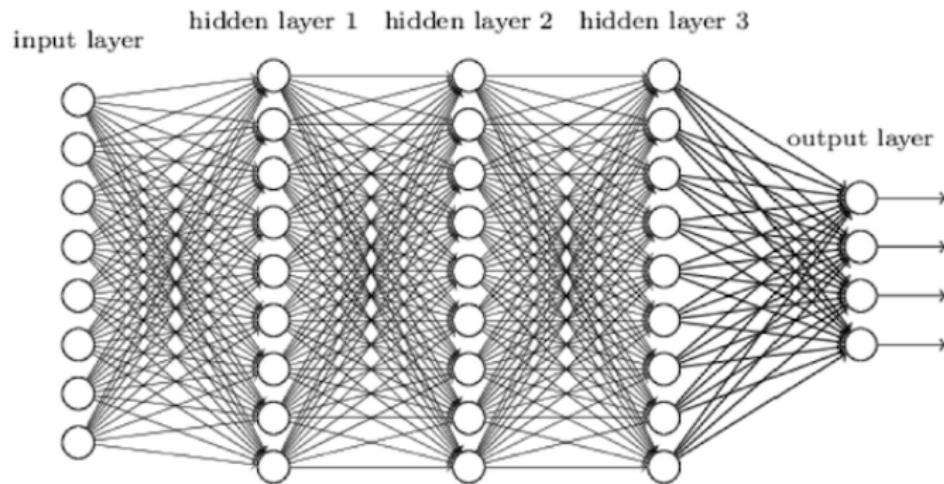
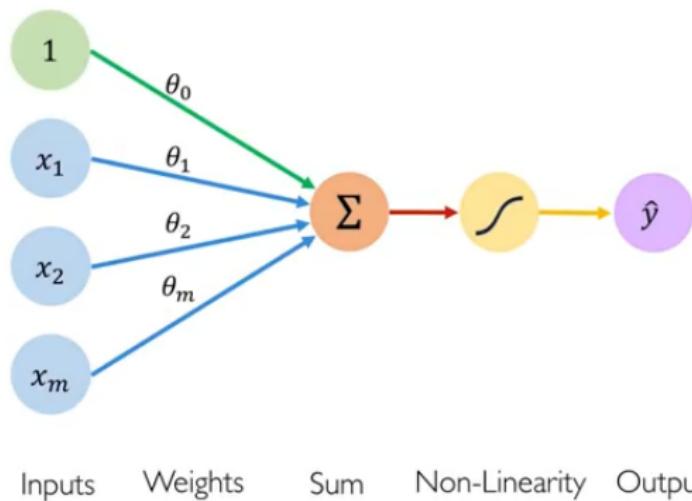


Figure: Deep Learning

Deep Neural Net



Linear combination of inputs

Output

$\hat{y} = g \left(\theta_0 + \sum_{i=1}^m x_i \theta_i \right)$

Non-linear activation function

Bias

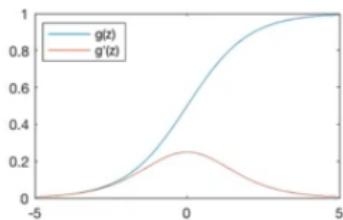
Diagram illustrating the mathematical formula for forward propagation. The output \hat{y} is calculated as the non-linear activation function g applied to the linear combination of inputs. The linear combination includes a bias term (θ_0) and weighted input terms ($x_i \theta_i$).

Figure: Forward Propagation

Source: MIT 6.S191 (2018) - Introduction to Deep Learning

Deep Neural Net

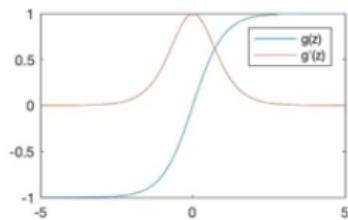
Sigmoid Function



$$g(z) = \frac{1}{1 + e^{-z}}$$

$$g'(z) = g(z)(1 - g(z))$$

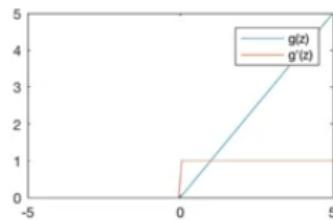
Hyperbolic Tangent



$$g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$

$$g'(z) = 1 - g(z)^2$$

Rectified Linear Unit (ReLU)

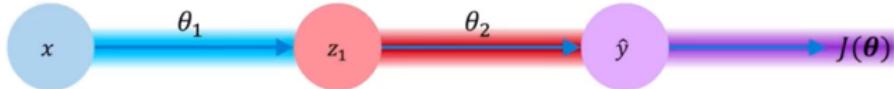


$$g(z) = \max(0, z)$$

$$g'(z) = \begin{cases} 1, & z > 0 \\ 0, & \text{otherwise} \end{cases}$$

Figure: Activation Functions

Back Propagation



$$\frac{\partial J(\theta)}{\partial \theta_1} = \underbrace{\frac{\partial J(\theta)}{\partial \hat{y}}}_{\text{purple}} * \underbrace{\frac{\partial \hat{y}}{\partial z_1}}_{\text{red}} * \underbrace{\frac{\partial z_1}{\partial \theta_1}}_{\text{blue}}$$

Figure: Loss Minimization

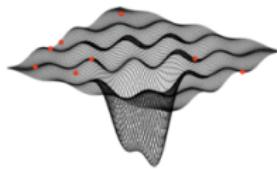
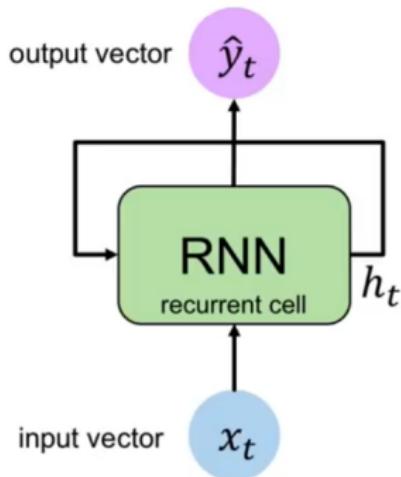


Figure: Global versus Local Minima

Repeat this for every weight in the network using gradients from later layers.

RNN



Apply a **recurrence relation** at every time step to process a sequence:

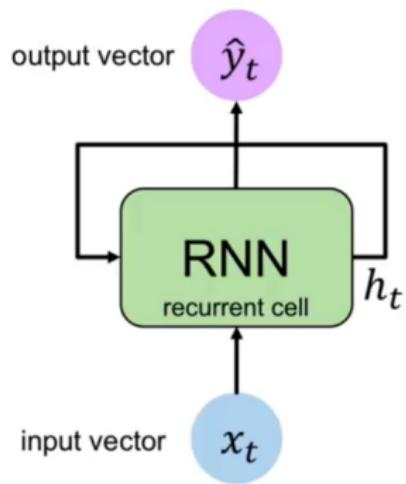
$$h_t = f_W(h_{t-1}, x_t)$$

cell state function parameterized by W
old state

Figure: Same function and set of parameters are used in every step

Source - MIT 6.S191- Recurrent Neural Networks

RNN



Output Vector

$$\hat{y}_t = W_{hy} h_t$$

Update Hidden State

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$$

Input Vector

Figure: State Update and Output

RNN

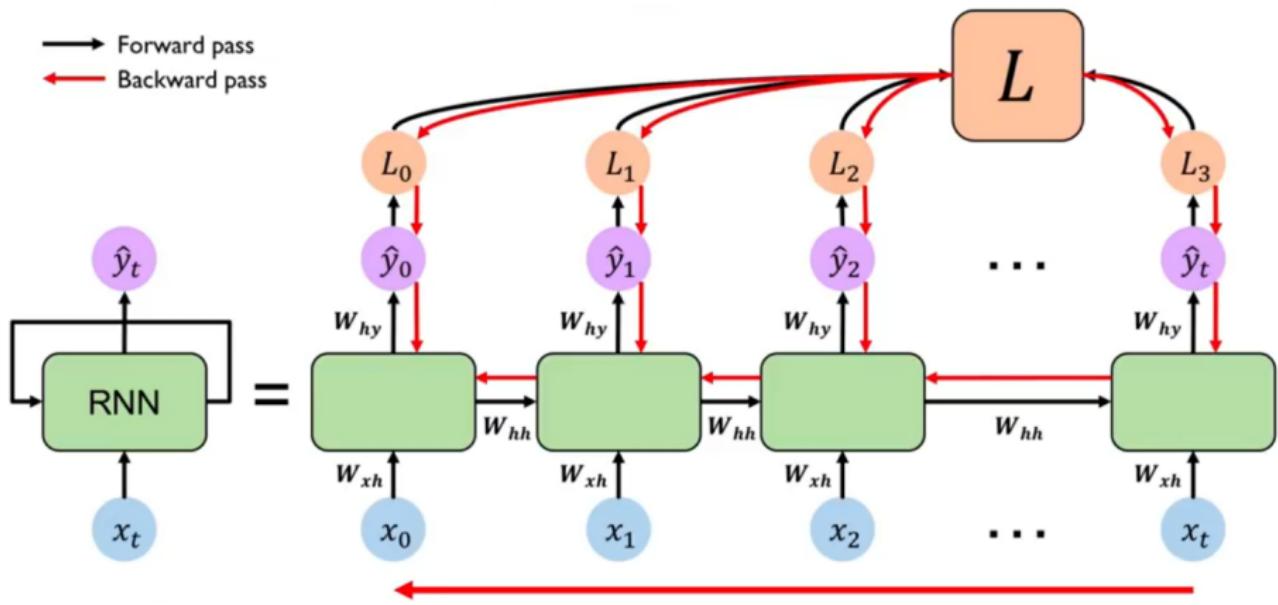
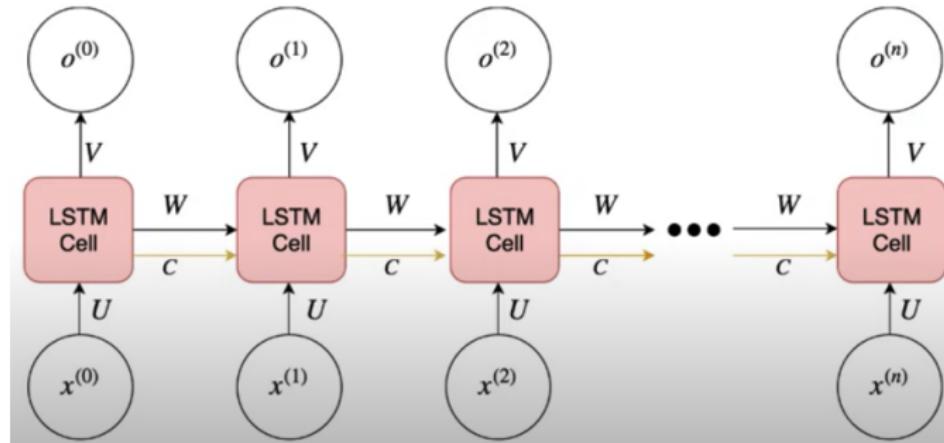


Figure: Backpropagation over time

Vanishing / Exploding Gradients

- ① Learning with recurrent networks is challenging due to the difficulty of learning long-range dependencies, [Bengio et al. 1994], [Hochreiter et al, 2001].
- ② Problems of vanishing and exploding gradients occur when backpropagating errors across many time steps.
- ③ Which of the two phenomena occurs depends on whether the weight of the recurrent edge $|w_{jj}| > 1$ or $|w_{jj}| < 1$ and on the activation function in the hidden node. For a sigmoid activation function, the vanishing gradient problem is more pressing, but with a rectified linear unit $\max(0, x)$, it is easier to imagine the exploding gradient.

Long Short Term Memory (LSTM)



Maintain separate cell state to avoid vanishing gradients by allowing the network to selectively remember or forget information over long periods of time by incorporating a set of gates that control the flow of information through the network.

$$i_t = \text{sigmoid}(W_i[x_t, h_{t-1}] + b_i)$$

$$f_t = \text{sigmoid}(W_f[x_t, h_{t-1}] + b_f)$$

$$C_t = f_t * C_{t-1} + i_t * \tanh(W_C[x_t, h_{t-1}] + b_C)$$

$$o_t = \text{sigmoid}(W_o[x_t, h_{t-1}] + b_o)$$

Questions?

Thank You!