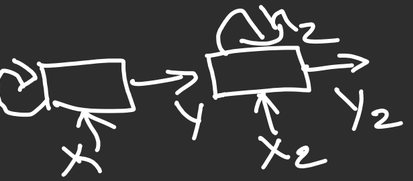
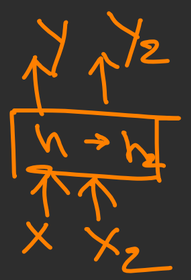
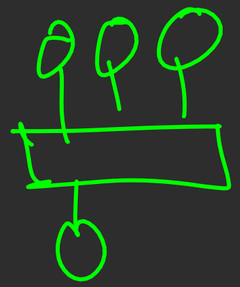
Rnn is serial → 

Transformer is Rnn for parallel computing →



Rnn used

many to one
ex: → sentiment



one to many
ex- image captioning

many-many
ex - Translation

one hot → is a type of embedding

Back propogating Through Time

feed forward wali algo hain

exploading gradiant me
bohot jaldi ... badhte hain
vanishing gradiant ulta hota
hain i.e ex (0.01 × 0.1 = 0.001)
gradiant clipping — exploading
to kam karne ke liye
vanishing ke liye tumhe starting
weight aur bias jisse se initialize
karo

vanishing se bachne ke liye
→ relu use karo kyo ke positive
deta hain
→ pananct adhe se rakho
→ use gate to view long term
dependencies matlab
dhoka yad rahe
(AKA LSTM) use karo

Transformer → it work on
attention mechanism

number of features → attention → useful features

necessary related word/ya
feature dekhte hain

self attention → Query, key
Q, K, V value

Q, K ke beech me similarity
with vector positioning
phir Q, K ki similarity ko
to function softmax me
dete hai phir V se
multiply karte hain

$$= \text{softmax}\left(\frac{Q \cdot K^T}{\text{scaling}}\right) \cdot V$$

Positional embedding x Linear layer

= Q

" x L_2 = K ← But different

" x L_3 = V