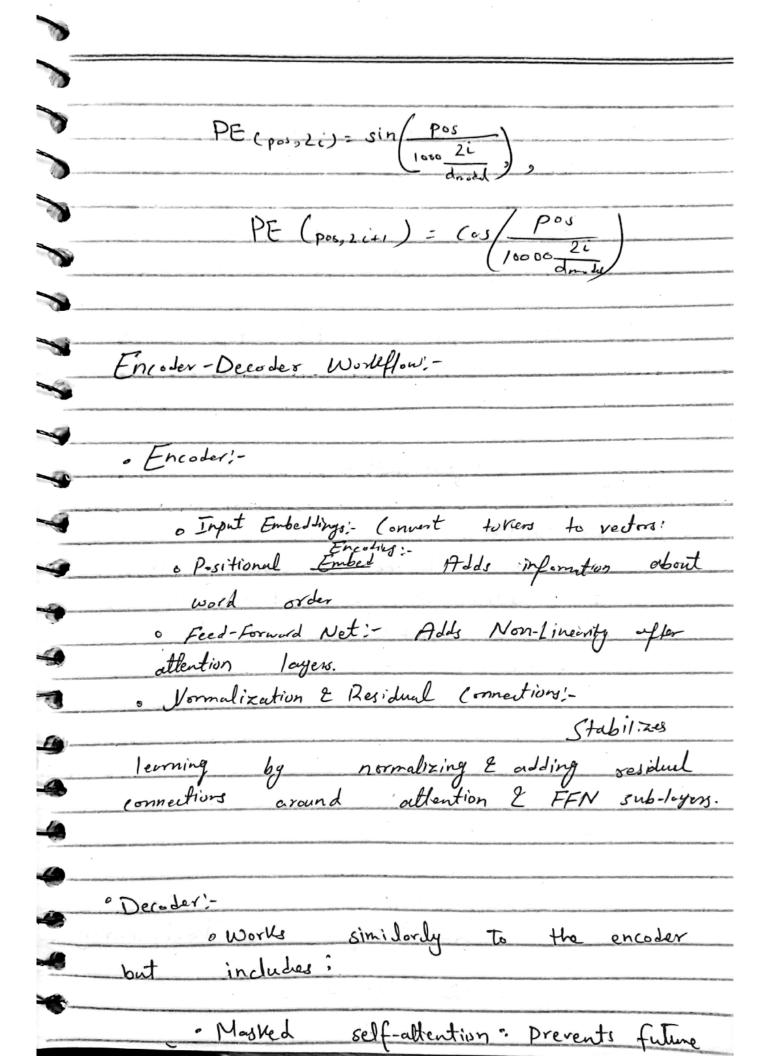
	1
Attention (Q. K.V) = softman (QK)	0
Adu J	The same of the sa
· Scaled Dot-Product Altention; - Normalizes the	200
dot product by dividing by ldk (dinension	SIN
of keys) to prevent large values.	7
	ATT.
Multi=Itend Attention!	
Instead of a single altertion layer,	100
in parallel. The results are concatenated:	
	1
Multi-Head (Q, N, V) = (oncut (head, , head, , heady) h	(A)
1 room (100 (40 h) - Contral (head), head, i-heady h	Jo D
whore each alterth head is:-	6
nee adie Attention (QWi oKWik, VWi)	
	*
	_*
Positimul Encoding:	
Since transformers deon't	1
proces sequences in order, positional info	(
is added'r	



	VIII THE REAL PROPERTY.
tokens from being altended to du training.	ning
o Encoder-Decoder Attention: Aligny decoder	, ulput
with the encoders's processed input	
is encountered.	
Trajaing & Regularizati	
1- Label Smobthing!-	
Instead of one-hat encoding, the target distribution is smoothed by adding some probability mass to in	
adding some probability moss to in labels , improving generalization,	
P'(i) (1-E) P(i) + V	
2- Loss Function:	
Typically uses cross-entropy 1. Incorporating KL Divergence for label -5. P/(i)/ay/2011	The second secon

S	
6	
9	
1	
)	3- Optimizer-
	Uses Adam optimizer w/
))	Terrning rule scheduling:
100	
1	
3	Irute (step) = dmotel . min (step , step. www.p. uty
-	
	Dalala CNINCI-
	Advantages Over RNNs/CNNs!-
4	DAIN JULY ANNUAL
-	Parallelism: Valike RNN's that process
	data sequentially transformers process all
4	tokens simultaneously a making training much
4	forster.
3	Long-Range Dependencies!
4	Long-Range Deper Transformers handle
	tokens be ther
3	ict and geral to see the
8	due to self-attention mechanism
4	
4	
	
•	