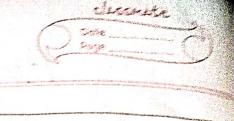
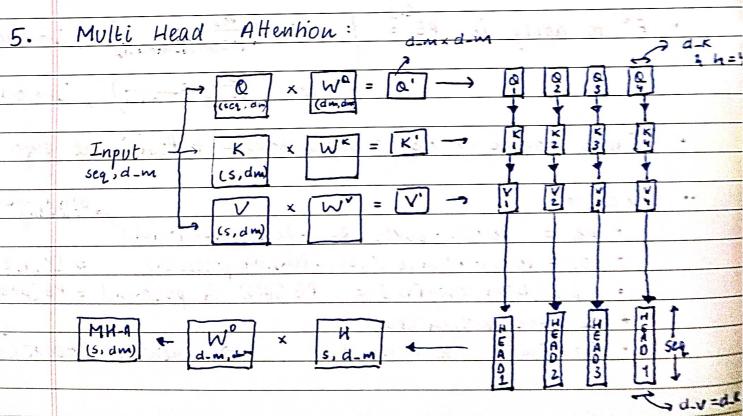
	The state of the s
	Attention Is All You Need
	1 Toolactou 13 ATT TOO NEED
	Transformer
I.	
	ARCHITECTURE =>
1.	Juput Embeddings =>  cat -> 1234 -> 1000  Unique Input 1d 6084  d_model
•	cat - 1234 - 1000 1
	Unique Input 1d
	324 8828
	the 11 to being the second of
•	Whole tent - seq-len x d-model
	Whole tend → seq - len x d-model  > ofm □ □ □ □ □ □ (seq-len=6)
••	in the state of th
-	Code →
الما و العمل بد	• IE Layen = mn. Embedding (vocab-size, d-model) • Find fn = IE(u) * Jd-model [1]
	+wa +n = IE(u) * Ja-model
	The state of the s
2.	Positional Encoding:
	V
	F" to apply it, PE = \( \frac{\text{Pos}}{10000\\ \alpha\rmodel} \right) \\ i\rightarrow\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \
5	
	cos (Pos (100.00 a model), i - odd)
•	Input token -> P.E - suput token + Positional sujo
••	*Coda ** 1
	· pe = toych · zevos (scq - leu; d-model)
S   .	• pos = torch·avange (0, seq-lu). unsqueeze (1) => [0,1,2s]
9	pe [even columns] = PE (pos) 4 pe [o.c.] = PE (pos)
1	* negister_buller (pe)  Fwd: fn = n + pe - nehvn: dropout (n)
1 . 1 . 1 . 1	Cal (a) [a) According to Decay
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	(11, [2], [3] => According to Paper



- Layen Normalisation:
- $\alpha = \text{torch} \cdot \text{ones} (\text{features})$
- atures),  $\beta = \text{torch. twos}(f)$   $\Rightarrow \alpha, \beta^{(4)}$  Leounable nn. Parameters ()  $\mathcal{E} = 10^{-9}$
- Fwd fu=> ni - meancus Jstd dev(xi) + E
  - This mean is wet dim=-1
- Feed Forward Network: 4.
  - Simple NN with dropout to train weights & biases

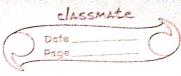
    FFN(x) = max (0, xW1+bc) · W2 + b2 [6]

  - Code FFN\_layers = nn. Seg (Linear (d\_m, d\_ff), Relu, dropout, Linear (d-ff, d-mode



[4],[5],[6] - Acc to Paper

\* 47



and the same of th	
	d-model = d.
	d-model = dm => dm 1/h = dk = dv
•	
	softman OK.
	Attention $(0, K, V) = softman (0 K^T) \cdot V$ $\sqrt{dk}$
•	
	· init => $d_{K} = d_{K} = d_$
	· init => dx = dv = dm 11h wq, wk, wv, wo = nn. Linear (dm, dm)
	Mo = MM. Lineau (dm, dm)
	* attention => att _ scores = ( query * key T) /Jak
	att_scores · masked_fill_ (0, 10-9)
	Hehren Soll and att 1/2/10
-	return softman (attusc): * Value
	* forward ≥ (q,k,v, mask) =>
	guery = w-q(q), key; value similarly (Batch, SL, d.
	query = query. ushape (q.shape[o], [1], h, d-k). transpose (1,2
1.0	# g=d=m - h d-k -, Final shape - (B, h, sh, d-k)
	111 4 key & value.
•	1114 key & value $\mathcal{H} = \text{aHention}^{\bullet} \left( \text{qvery, key in ask} \right)$
B,SL,du	) = x = x. transpose (1,2). contiguous (). view (x.shape[0],-1, d.
-	return wo(u) will sold reposed
1	should be Ingot Enough + Par Encoding + T
6.	Residual Connection:
1	
4 46 6	2 Layer + 20 2+2,
	the second second
1	: Code => netwer x + (dropo ut (sublayer (Layer Norm (n)
4 40	Code => netween x + (dropo vt (sublayer (Layer Norm (n) Layer being applied

CAJ - ATP

	Date Page
Se	Encodeu:
Na Pi	
yer	Encoder Block =>
48	init -  ( featives, attent - block, ++n, avopous)
-	of a black = ath block teed-two-net = ffn
ar A	medicinal couns = ModuleList (RN(2))
moto	
inger!	- fwd = ( m, s.vc = mask)  - fwd = ( m, s.vc = mask)
	OH OUTOUT = UTEN - DIOCK (X)
	N = residual - couns (o) ( 2, att - ourpour
	n= residual - conns [1] (n, ffu).
	neturn kind.
	Encoder =>
	initi=>: layers_in-encoder ; Layer Norm
-	man from the first telephone of the first of
	fud > for layer in Lize > . n = layer (n, mask)
	netwen LN(n)
and the same of the same of	( 21 Solver, programme and a solver of the
-	( ) ( ) ( ) ( ) ( ) ( ) ( ) ( ) ( ) ( )
-	Encoder Block Diagram =>
-	
-	Input - Input Embed - Posu Encoding + I I
	i divide a la faction de la fa
-	I PIN A
-	XN- <= 1 Act of a Norm

1 mark

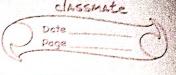
netwin LN.(n) Encoder Block Diagram =>

7.

Input - Input Embed - Posu Ence .... xN. Encoder ( deposit (supply of the )

[8], diagram - ATP

0.	Decoder:
0.	Decoder:
*	Decoder Rioch ->
	Decoder Block =>
1 1	init (self_att_block, cross_att_bl, ffu, dropout) =>
	sab, cab, Ifu initialish [9]
-	res_conns = Module List (Res Con (3))
16. 1	fwd (u, encoder_output, tgt=mask, syc-mask):
	x = ues_con[o] (x, att n output)
	X= Mes-con[1] (x, cross-aH output(x, eo, e0, mask))
	x = eus - con(2)(x, ffn)
10 100	Helbru ni - man de la company
	- Ne coder and service of the servic
* )	-Decoder = 1 30 minum e 29 e 21 e 424
	init - layers, layerNorm
	fud > for layer in layers > x = layer(x, e0, src_mask, +_n)
2 2	Heturn self-norm(n)
ال المنطق ال	Decoder Block Diagram =)
, J. E 1	a me on the properties of the state of the s
* 1	DAN 17
	TOTAL STATE OF THE PERSON OF T
	ALN
	CYOSS MHA
791	
	output ( ) X N
The X	Action - second to Add & Norm
1. 10.	
- 4	with mast
	and which is an most spare
	STILL TO THE TENED VII ALL TONG
(197	Output Embeddings   [9], Diag -> ATP
	Outputs ATP
	Outputs who till the

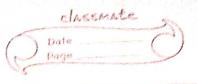


g. Projection Layer => · init => proj = nn. Linear (d.m., vocab-size) fwd => neturn log\_softman (\* proj(u), dim=-1) TRANSFORMER => Creqtes structure for Transformen. 10. init => encoder, deoder, src-embed, tgt\_embed, src-pos tge-pos projulayer encode = (src, src\_mask) = src - Input Emb -> Posm Enc -> Masking & Encoder Fm decode (encoder-off, src-mask, tgt-mask, tgt - IE -> PE -> Masking svc & tgt & Decoder F" · project (n) => project (u) BUILDING TRANSFORMER => Juitialises Transformer 11. build transformen (src\_vocab\_len, t\_v-l, src-seq-len, t-s-l, d-model = 512 N=6, h=8 dropout = 0.1, d-ff = 2048 C10] => src\_embed, tgt\_embed from src, tgt\_vocab-size src-pose, tgt-pos from src-seq-len 2 +-5-l encoder-blocks = [] - append encoder block after init décoder-block = CJ - append décoder block after int encoder from e-b & clecoder from d-b proj\_layer using tgt\_vocab\_size. transformer initial" & setting parameters (xavier)

[10] - ATP data.

$\pi$	DATA
	at a Vikeraty; 110
12.	Bilingual Dataset:
•	init ( ds, tokenizur syc. + + cri land + 1 co. 1)
W	init ( ds, tokenizur svc, +-t, src-lang, +-1, seq - len) =>
	initialise all these 1 +
	505 - token eac tolers
	50s - token, eos - token, pad-token using token-to-id
	len
	- getitem (idx)? and I whom it whom it is
	Src-tgt-pair = ds[iadeu] -> src-teut, t-t
	enc_input_token, d_i_t 2
F1	enc h dec pad tokens.
	encoder_ip = [sos, enc_input, pad = enc]
	III'y dec helabel.
13.	Build Tokenizer, Ogtaloaders h Transpormers =>
•	get_all_sentences - yield all items
	· ·
•	get_or_build_tokenizer(config , ds , larg) =>
	Check tokenizer else
	tobenizer = sub-word Iword level
	-
•	get - ds (config) =>
	de load, get suc & tgt to kevi een, train_test split
	train-ds, test-ds, man-leu-sve h man-leu-tgt
* 1.75	building train, test dataloaders
	C++3
	get_model() => model = build-transformer ()

E113 + 11.44's fm.



and the second s	
I.	DATA
12.	Bilingual Dataset:
	12/10/10/10/10/10/10/10/10/10/10/10/10/10/
•	Class:
. 4	init (ds, tokenizur svc, +-t, src-lang, +-1, seq-len) =>
	initialise all these 1 +
	50s - token, eos - token pad-token using token-to-id
	S. Williams
	len + return len (ds)
	- getitem (idx)?
	src-tgt-pair = ds[iadeu] - src-teut, t-t
	enc_input_token, d-i-t-2
9 =-	enc l' dec pad tokeus.
	encoder_ip = [sos, enc_input, pad = enc]
	1117 dec 4 label.
13.	Build Tokenizer, DataLoaders h Transformers =>
•	get_all_sentences -> yield all items
•	get_or_build_tokenizer(config , ds , larg) =>
	Check pokenizer else
200	tobenizer = sub-word Iword level
•	get - as (config) =>
	de load, get suc & tgt to benizer, train-test split
и п	train-ds, test-ds, man-leu-src h man-leu-tgt
	building train, test dataloaders
	CH3
	get_model() => model = build-transformer ()
Fac. 2	

E11] a 11. m's f".

THE K	
	Date
	Page Page
Ⅲ.	TRAINING =>
14.	Con Lig File =>
	act cowlig => returns config
	config File =>  get -config => networks config  get wt file path => Makes file path for each epoch latest - weights - file - path => Finds who of newst save
	Solent superplates file natha => Finds sults of sales int
	Tarest _ to agras _ TITE _ parts
	*
· , # ; ; .	to the section and the section of th
15.	Training =>
	Initialises everything loss, ophin, path device
	Initialises energhting loss, optim, path, device.  Loads model if samed model present.
	Training Inpo
	Training loop.  Saving Model
	Journey Journal of the Control of th
1.5	
16-	Evaluation =>
	Evaluates the 1651 saved, model:
	a service of the serv
	wasting their a second to the second
	The state of the s