

St-Moe

Barret Zoph^{*}
Google

Irwan Bello[†]
Google Brain

Sameer Kumar
Google

Nan Du
Google Brain

Yanping Huang
Google Brain

Jeff Dean
Google Research

Noam Shazeer[‡]
Google Brain

William Fedus
Google Brain

ABSTRACT

Experts MOE -

269b complacaleto A B
- Experts ST-MOE- B

Superglue Arc Easy ARC XSUM CNN-
DM WebQA

Winogrande Anli R 1

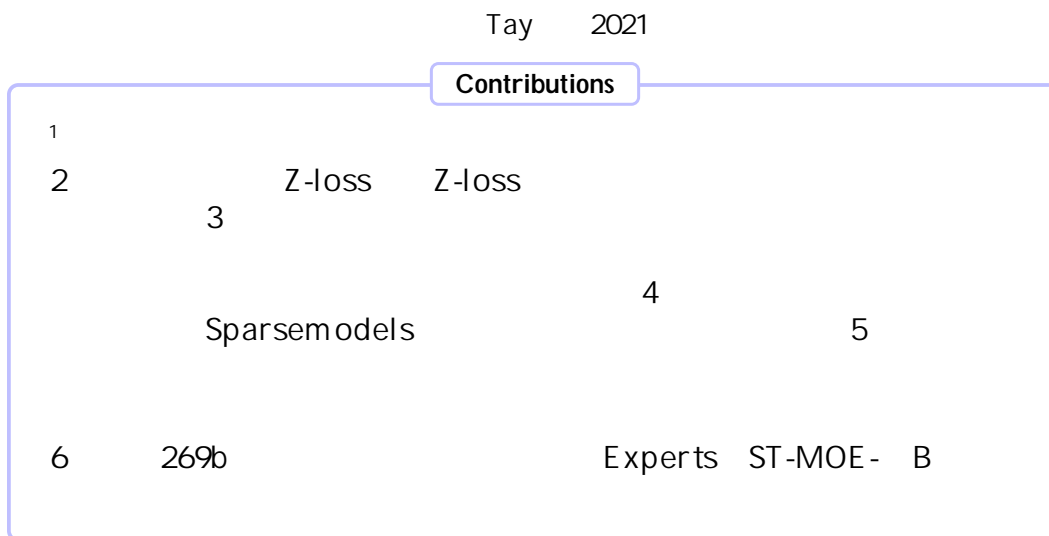
^{*} {Barretzoph, liam.fedus}@google.com [†]
Google. for https://github.com/tensorflow/mesh/mesh/master/master/master/mesh_tensorflow/moe/transformer/moe_py [‡]

CONTENTS

1		3
2		3
3	3.1	5
	.3.2	6
	.3.3	6
	.3.4	7
		8
4	4.1	9
	.4.2	9
	.4.3	11
	.4.4	11
		12
		13
5	5.1	13
	.5.2	13
		14
6	Experimental Results	16
6.1	ST-MoE-L	16
6.2	ST-MoE-32B	16
7	Model .	19
	.7.2	19
	.7.3	19
		21
8		21
9		22
10		24
		31
B	Z-loss	31
c		32
D		33
e		34
f		35
g		36
	H	36
		37
J		38

1

Expert networks
Rae et al. 2019; Brown 2020; Rae 2021
Shazeer 2017
Lepikhin et al. 2020
Fedus 2021; Artetxe 2021; GPT-1/3
Du 2021
Patterson 2021
Fedus 2021
Rae et al. 2019
Artetxe 2021
1.6T
Oncommon 4
Supersglue Moe 8
FLOPS
Th Thrage T
Switch-XXL
Du 2021
269B
Superglue
SuperGlue
NLP



2

Jacobs 1991 Jordan Jacobs 1994 have unique
2021; Roller 2021; Zuo et al. 2021; Clark 2022 Lewis

Shazeer 2017 Experts MOE
 $\{E_i\}_{i=1}^K$
 $N_i = 1$
 $N = \sum_{i=1}^K N_i$
 W_r
 $\text{logit}_i = w_r \cdot x$
 $p_i(x) = \frac{e^{\text{logit}_i}}{\sum_{j=1}^K e^{\text{logit}_j}}$ (1)
 $y = \sum_{i=1}^K p_i(x) E_i(x)$ (2)
 Shazeer 2017 LSTM Hochreiter 1997 Vaswani 2021
 MOE Lepikhin 2018 Schmidhuber 2020 Fedus TOP-1

Terminology	Definition
Expert	An independently-learned neural network with unique weights.
Router	A network that computes the probability of each token getting sent to each expert.
Top-n Routing	Routing algorithm where each token is routed to n experts.
Load Balancing Loss	An auxiliary (aux) loss to encourage each group of tokens to evenly distribute across experts.
Group Size	The global batch size is split into smaller groups, each of size Group Size. Each group is considered separately for load balancing across experts. Increasing it increases memory, computation, and communication.
Capacity Factor (CF)	Each expert can only process up to a fixed number of tokens, which is often set by evenly dividing across experts, $\frac{\text{tokens}}{\text{experts}}$. The capacity factor can expand or contract this amount to CF $\cdot \frac{\text{tokens}}{\text{experts}}$.
FFN	Acronym of Feed Forward Network (FFN) layer of Transformer consisting of linear, activation, linear.
Encoder-Decoder	A Transformer architectural variant that all of our models are based on. Consists of an encoder that does all-to-all attention on the inputs and a decoder that attends to the encoder and to its own inputs in an autoregressive manner.
all reduce	Communication primitive which sums a subset of n tensors on n different devices, then broadcasts the summed value to all n devices. This is used in distributed training for gradient accumulation and model parallelism.
all2all	Communication primitive where each device sends to every other device a part of its tensor. Used in sparse Transformer models for token routing.
(\uparrow/\downarrow)	Indicates whether higher/lower values are better (e.g. accuracy/train loss).

1
 B 2 b/g CF CF /
 2 /fromexperts Einsums

2017
The loss
Clark 2022

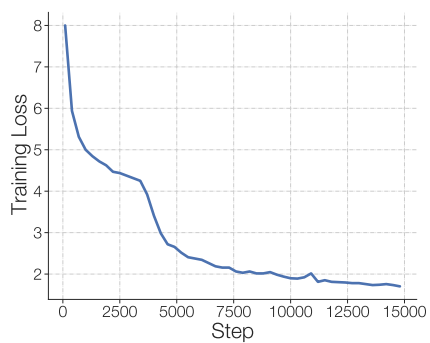
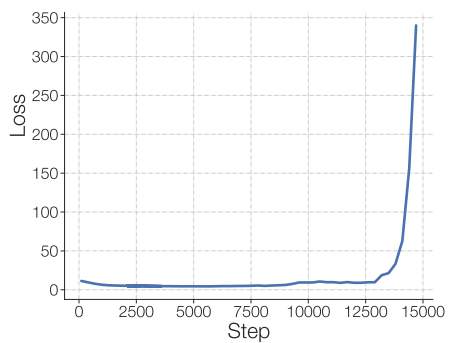
A

3

Shazeer
Lewis 2021

3

1 Stan-Dard



1

T-XL
1M Shazeer and Stern 2018

Ra el 2019

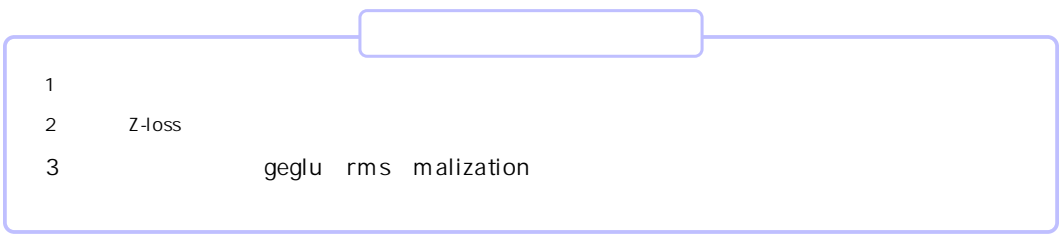
Afaactor Optimizer

1 Sta-bility

2 Z-loss

3 Z-loss

Shazeer 2018



Xue 2020

T-XL 32 Ra el 2019

MUE

3 Fedus 2021

FFN

1.25

2.0

11

theStandard

T -XL
Butwas

1/3

9

Compute

2018

3.1

" "

GELU
al 2017
Shazeer
Gimpel 2016 FFN

Dauphin et al. 2010
Sigmoid
Aggelos et al. 2010
Nair and Hinton 2010
FFN

$$FFN_{GEGLU}(x; W; V; b; c) = GELU(xW + b) \quad (xV + c) \quad (3)$$

Narang 2021

Root MeansQuare RMS

Zhang Sennrich 2019
Sublayer

gragradien t

Sublayer
2

Ra el

2019

1 RMS

2014

He

2015

RMS

 $x \in R$

$$G.y_i = x_i \left(\frac{1}{d} \sum_{i=1}^d x_i^2 \right)$$

2

gegl u

RMS

g

FFN

Shleifer et al 2021

C

3.2

Taleb

Nee-Lakan tan 2015

Method	Fraction Stable	Quality (‘)	
Baseline	4=6	-1.755	0.02
Remove GEGLU	3=3	-1.849	0.02
Remove RMS Norm. Scale Param	3=3	-2.020	0.06

2

geglulayer

3 jitter 2021 XL Fedus
-jitter [1-10-2 1 + 10-2]
Kaplan 2020

Method	Fraction Stable	Quality (‘)	
Baseline	4=6	-1.755	0.02
Input jitter (10 ⁻²)	3=3	-1.777	0.03
Dropout (0.1)	3=3	-1.822	0.11

3 jitter TOA

3.3

Szegedy 2015; Salimans and Kingma 2016; Ba Pascanu 2013; Iosif and
Pascanu 2013
Afaactor 8 Dettmers 2021
Dettmers

Fedus 2021 Float
Z-loss

$$L_z(x) = \frac{1}{B} \sum_{i=1}^B \sum_{j=1}^N \log \frac{e^{x_j^{(i)}}}{A} \tag{5}$$

b n x Rb x n logits
3.4 Z-loss

4 Z-loss 3 Z-loss Theupdate 4

Method	Fraction Stable	Quality (")
Baseline	4=6	-1.755 0:02
Update clipping (clip = 0:1)	3=3	-4.206 0:17
Router Z-Loss	3=3	-1.741 0:02

4 inadafactor

Z-loss

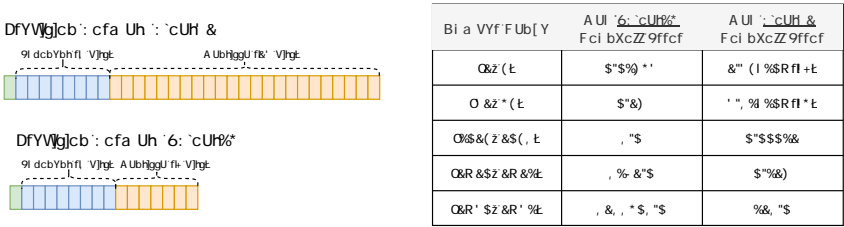
Z-loss C Z
L CE L B Z-LOSS L Z

$$L_{tot} = L_{CE} + c_B L_B + c_Z L_Z \tag{6}$$

PA C Z = 0.001 B

3.4

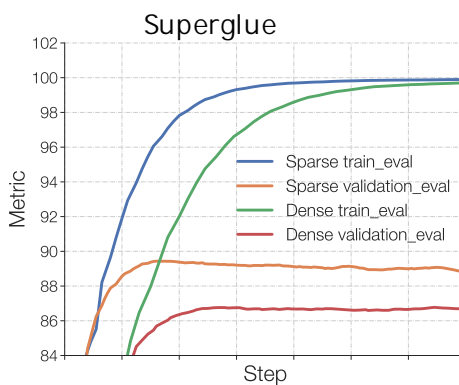
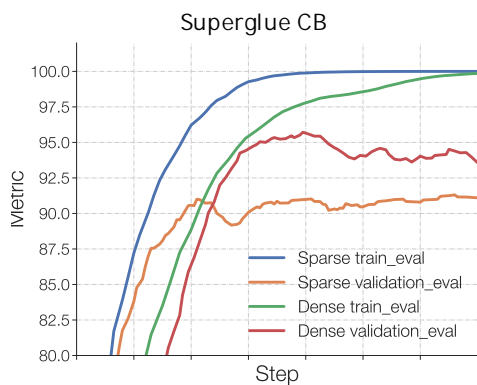
oat Micikevicius 2017 5
B oat B oat BFLOAT
Float ActivationSare
ST-MOE- B Alleduce
Allreduce
Float
a b
c Activa-tions



2 Float 65,536 z-lossenciring BFLOAT
Z-loss

4 Z-losses 5 lity
https //github.com/tensor ow/mesh/mesh/
blob/blob/master/master/master/mesh_tensor ow/mesh_tensor ow/ Matrix h/
6 Matrix 6 Matrix TPU BFLOAT Float 2.

2
 Mantissa Finumber B oat 2 [2,4 [1024 2048
 65,536X 23 - 7 = 16 7 Float 23 1 B oat
 As oat 2 2 16 = 65536 8 NumberCan
 3E Float
 - - 7
 MOE /5 asthat
 -
 9
 Z-loss Z-loss logits
 Wu 2016 logits ROUND OFF - ValueAnd
 Z-loss oat
 Z-posses
 9
 4
 1 ofdata
 2 Superglue
 Lester 2021 Brown 2020 Houlsby Li and Liang 2021
 2019 - Du
 2021 ; Artetxe 2021
 Ouyang 2022
 4.1
 Fedus
 2021 Artetxe 2021
 Superglue Wang 2019 - De
 Marne eetal 2019 Zhang 2018
 CB 250 100,000
 3 L ST-MOE-L C 500B
 Ra el 2019
 7
 128 logitwith a 128.5 SoftMax 10
 SoftMax 36 BFLOAT 0.5
 Exp + ·Exp 0.091 SoftMax
 128.5 128 B oat B oat
 situsution oat logit



3 CB
ST-MOE-L
Eversgreen Line
The-Ex-Out Red vs. Orange
Outperform

250 138K
Blue
CB

2019 770m
FFN MOE
FromRael
ST-MOE

T-Large
32 1/4 2.0
1.25

100

The dense

Fedus

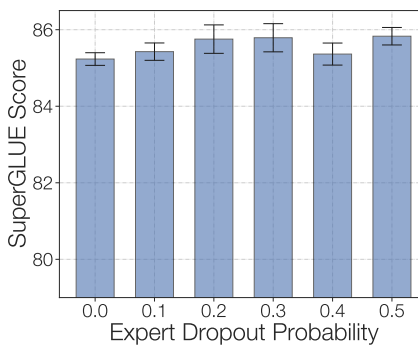
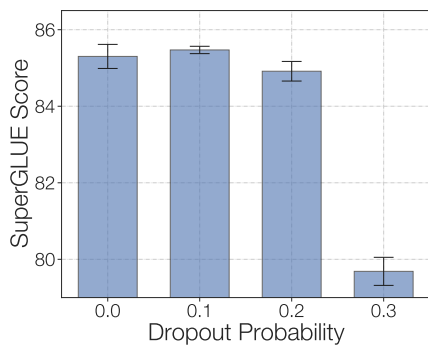
2021

CB

4

4.2

4.3



4

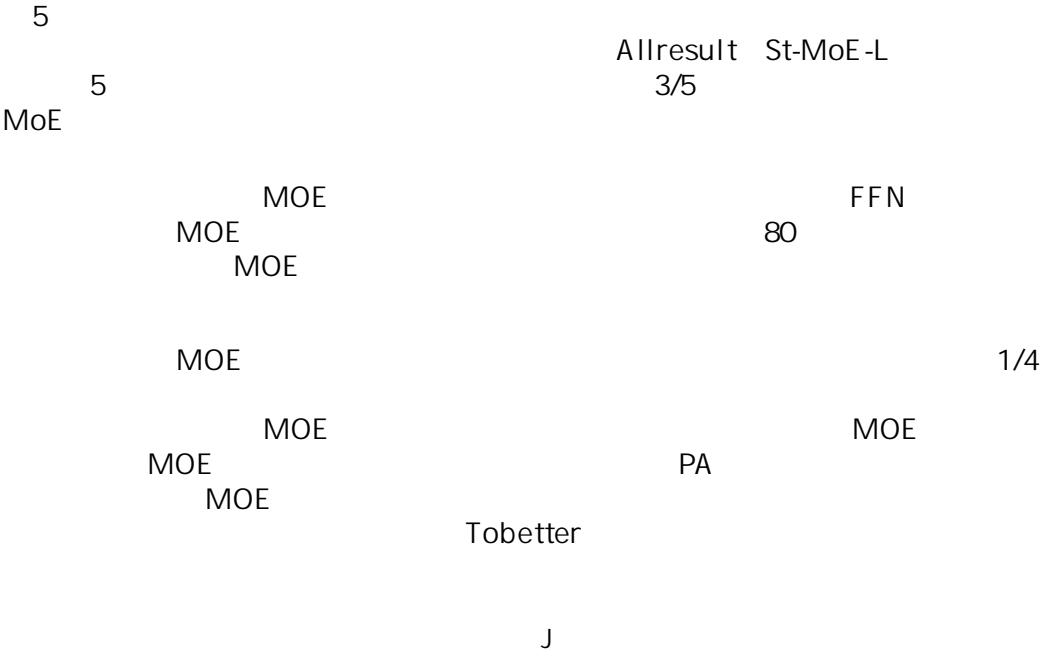
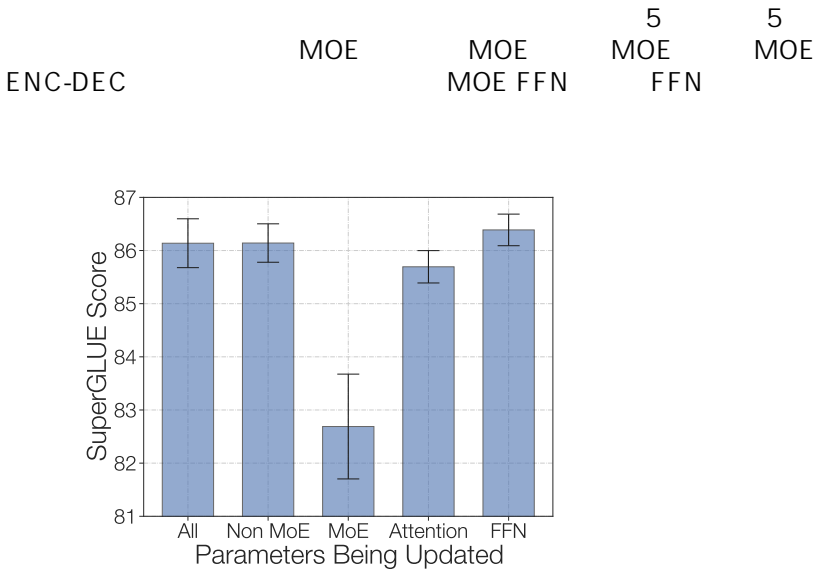
0.1 aglobal

0.1

Fedus

2021

4.2



4.3

500b Tokens of C

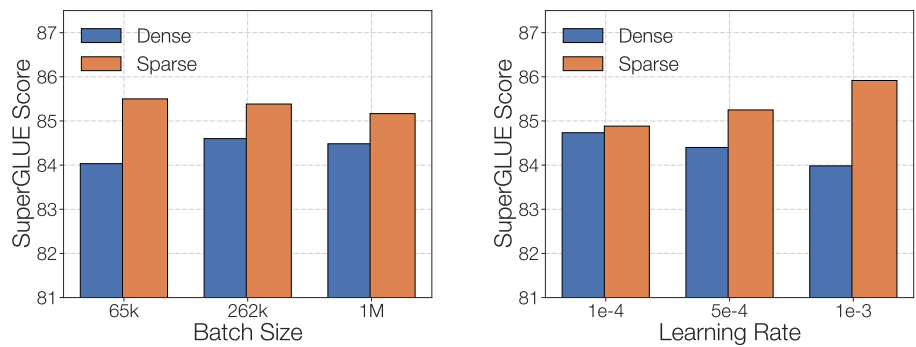
lenth-l 20

st-moe-l F

6

-

4.1



6
Superglue
6
Noisierhyperparameter

F

4.4

SPMD
1
2
- Thosetokens 2

5
ST-MOE - BCOROBATER
Yang 2021
8 8 10-15

Model	Train CF	Eval CF	Aux Loss	Percent Tokens Dropped	SuperGLUE (%)
Sparse	0.75	2.0	Yes	10.6%	86.5 0.21
Sparse	1.25	2.0	Yes	0.3%	86.7
Sparse	2.0	3.0	Yes	0.0%	85.8
Sparse	4.0	5.0	Yes	0.0%	86.4
Sparse	0.75	2.0	No	15.6%	85.7
Sparse	1.25	2.0	No	2.9%	85.8
Sparse	2.0	3.0	No	0.4%	85.9
Sparse	4.0	5.0	No	0.0%	86.4

5
10-15
AUX
<1

8
ion

Fedus 2018 Devlin et al 2018
6

6 wehighhighlight

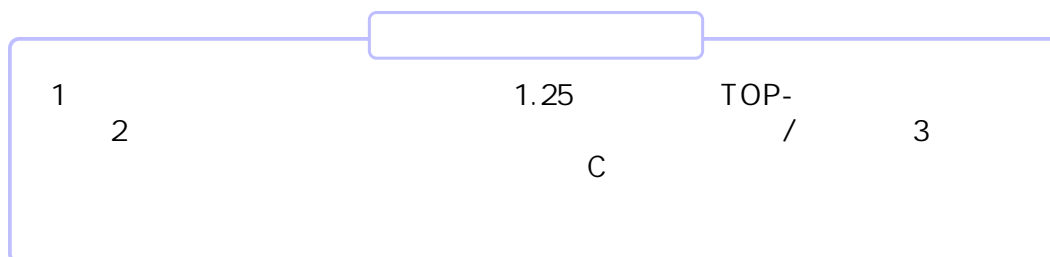
7
2021 Superglue GEC Rothe

Model	Insert Sentinel Tokens	SuperGLUE (")		GEC (")	
Dense		84.9	0.33	22.3	0.25
Dense	×	85.1	0.25	22.1	0.42
Sparse		86.6	0.18	22.2	0.04
Sparse	×	86.6	0.24	22.9	0.09

7 Weconsider ITDO Superglue Sentinel Lester 2021 GEC

```

graph LR
    Kaplan --- 1
    1 --- 2
    2 --- 2020
    2020 --- Sparse
    Sparse --- 3
    3 --- Fedus
    Fedus --- 2022
    2022 --- Clark
    Clark --- TOP-2021
    TOP-2021 --- 4
    4 --- Kaplan
  
```



Fedus 2021

Algorithm	Train CF	Eval CF	Neg. Log Perp. (")
Dense-L	—	—	-1.474
Dense-XL	—	—	-1.384
Top-1	0.75	0.75	-1.428
Top-1	0.75	2.0	-1.404
Top-2	0.75	0.75	-1.424
Top-2	0.75	2.0	-1.402
Top-1	1.0	1.0	-1.397
Top-1	1.0	2.0	-1.384
Top-2	1.0	1.0	-1.392
Top-2	1.0	2.0	-1.378
Top-1	1.25	1.25	-1.378
Top-1	1.25	2.0	-1.373
Top-2	1.25	1.25	-1.375
Top-2	1.25	2.0	-1.369
Top-2	2.0	2.0	-1.360
Top-2	2.0	3.0	-1.359
Top-3	2.0	2.0	-1.360
Top-3	2.0	3.0	-1.356

8

CF

EDARCF

CF

N RoutingAcross

N + 1

CF

Pareto

All

All

All

educe

N

TOP-N

N

All

All

/

All

reducecommunications

TPU

9

St-Moe-Land

ST-MOE-

B

ST-MOE-L

AllreduceCommunications

Fedus

ST-MOE-

B

2021

Lepikhin

2020; Du

2021

2.0

1.25

Model	Train CF	Step Time (s) (#)
ST-MoE-L	1.25	2.397
ST-MoE-L	2.0	2.447 (+7%)
ST-MoE-32B	1.25	4.244
ST-MoE-32B	2.0	4.819 (+14%)

9

TPU

1B

1.25

2.0

increas

32B

+7

theodel

14

ST-

MOE-L

8

St-MOE-

B

Top-N

J

Riquelme

D

BPR

2021

BPR All All Allreduce

6

T -Large 32B Ra el 2019 NLP 269B op

Wang 2019

SST- Word Sense Disambiguation WIC MRPC STS-B QQP

WIC sense session disambiguation WIC

MNLI QNLI RTE CB Question-words-ing Multirc Record Boolq

Coreference WNLI WSC COPA COPA

COLA NLP Superglue Toco correlate

CNN-DM Hermann

2015 BBC Xsum Narayan 2018

Rajpurkar 2016

Clark 2018 2020

Trivia QA Joshi et al AI 2017 2019 Web Berant 2013

Challenge Sakaguchi 2020 Winograndeschema

Nie 2019 NLI

6.1 ST-MoE-L

Ra el

UP TORMOM MAX NUM = 65536 2019

65536 Max NUM

10 T -LARGE L 500K

FLOP C DataSet 1M 524B

Ra el Ra el 2019 512 114

DEV Feduset

AI 2021 Roberts 2020

4.1 CB WSC

250 259 Sparsemodel

6.2 ST-MoE-32B

T -Large ST-MOE- B

2021 cododer-decoder Fedus

z-loss afafactor Optimizer

11 D F F 11 D KV 128

Fewer Switch-C Switch-XXL

11 Allreduce

he

Name	Metric	Split	Dense-L (")	ST-MoE-L (")	Gain (%)
SQuADv2	F1	dev	94.0	94.5	+1%
SQuADv2	acc	dev	87.6	88.1	+1%
SuperGLUE	avg	dev	85.1	87.4	+3%
BoolQ	acc	dev	87.1	88.6	+2%
Copa	acc	dev	83.0	91.0	+10%
RTE	acc	dev	91.0	92.1	+1%
WiC	acc	dev	70.4	74.0	+5%
MultiRC	F1	dev	83.9	86.0	+3%
WSC	acc	dev	95.2	93.3	2%
ReCoRD	acc	dev	85.7	88.9	+4%
CB	acc	dev	100	98.2	2%
XSum	ROUGE-2	dev	19.9	21.8	+10%
CNN-DM	ROUGE-2	dev	20.3	20.7	+2%
WinoGrande (XL)	acc	dev	75.4	81.7	+8%
ANLI (R3)	acc	dev	54.3	57.3	+6%
ARC-Easy	acc	dev	63.5	75.4	+19%
ARC-Challenge	acc	dev	50.2	56.9	+13%
Closed Book TriviaQA	acc	dev	28.1	33.8	+20%
Closed Book NatQA	acc	dev	27.2	29.5	+8%
Closed Book WebQA	acc	dev	30.5	33.2	+9%

10 op

250 CB 259

WSC

ST-MOE - B " " 269b 32B

C Switch-XXL In appendix C

1.5t C Ra el 2019

Glam Du 2021 Thedataset E Afactor

10K

Infedus 2021

12 Inference-

ST-MOE - B Superglue

91.2 93.2 Validation Accorsy

Forboth XSUM CNN-DM

Ra el 2019, Liang 2021 ST-

MOE - BIMPROV ARC Easy . . ARC 81.4 86.5

WebQA 47.4 42.8 Frouberts 2020

Ernie 3.0 Titan B Wang

2021 NATQA 41.9 Karpukhin 2020 41.5

Anli R Nie 2019 thestate thestate 74.7 53.4

ST-MOE - B SmallSquad

90.8 T -XXL 91.3

Model	Parameters	FLOPs/seq	d_{model}	FFN _{GEGLU}	d_{FF}	d_{kv}
Dense-L	0.8B	645B	1024	×	2816	64
T5-XXL	11.1B	6.3T	4096	×	10240	64
Switch-XXL	395B	6.3T	4096	×	10240	64
Switch-C	1571B	890B	2080		6144	64
ST-MoE-L	4.1B	645B	1024	×	2816	64
ST-MoE-32B	269B	20.2T	5120	×	20480	128
Model	Num. Heads	Num. Layers	Num. Experts	Expert Layer Freq.	Sparse-Dense	
Dense-L	16	27	–	–		
T5-XXL	64	24	–	–		
Switch-XXL	64	24	64	1=4		
Switch-C	32	15	2048	1=1		
ST-MoE-L	16	27	32	1=4	×	
ST-MoE-32B	64	27	64	1=4	×	

Name	Metric	Split	Previous Best (")			Ours (")
			Zero-Shot	One-Shot	Fine-Tune	Fine-Tune
SQuADv2	F1	dev	68.3 ^e	70.0 ^e	96.2 ^a	96.3
SQuADv2	acc	dev	62.1 ^e	64.6 ^e	91.3^a	90.8
SuperGLUE	avg	test	–	–	90.9	91.2
BoolQ	acc	dev/test	83.0 ^e	82.8 ^e	92.0	92.4
Copa	acc	dev/test	91.0 ^d	92.0 ^e	98.2	99.2
RTE	acc	dev/test	68.8 ^e	71.5 ^e	94.1	93.5
WiC	acc	dev/test	50.5 ^e	52.7 ^e	77.9	77.7
MultiRC	F1	dev/test	72.9 ^d	72.9 ^d	88.6	89.6
WSC	acc	dev/test	84.9 ^e	83.9 ^e	97.3	96.6
ReCoRD	acc	dev/test	90.3 ^e	90.8 ^e	96.4	95.1
CB	acc	dev/test	46.4 ^d	73.2 ^e	99.2	98.0
XSum	ROUGE-2	test	–	–	24.6 ^h	27.1
CNN-DM	ROUGE-2	test	–	–	21.6 ^a	21.7
WinoGrande XL	acc	dev	73.4 ^e	73.2 ^d	–	96.1
ANLI R3	acc	test	40.9 ^e	40.8 ^e	53.4	74.7
ARC-Easy	acc	test	71.9 ^e	76.6 ^e	92.7 ^g	95.2
ARC-Challenge	acc	test	51.4	53.2	81.4 ^g	86.5
CB TriviaQA	em	dev	68.0 ^e	74.8^e	61.6 ^b	62.3
CB NatQA	em	test	21.5 ^e	23.9 ^e	41.5 ^c	41.9
CB WebQA	em	test	38.0 ^f	25.3	42.8 ^b	47.4

2021 / Gopher Rae GPT-2021 Brown et al. 2020 Glam Du

2020 Weconsis C Ra el 2019 MC Xue
ST-MOE-L Themodel
32

< ID 0> _ < ID 1> _

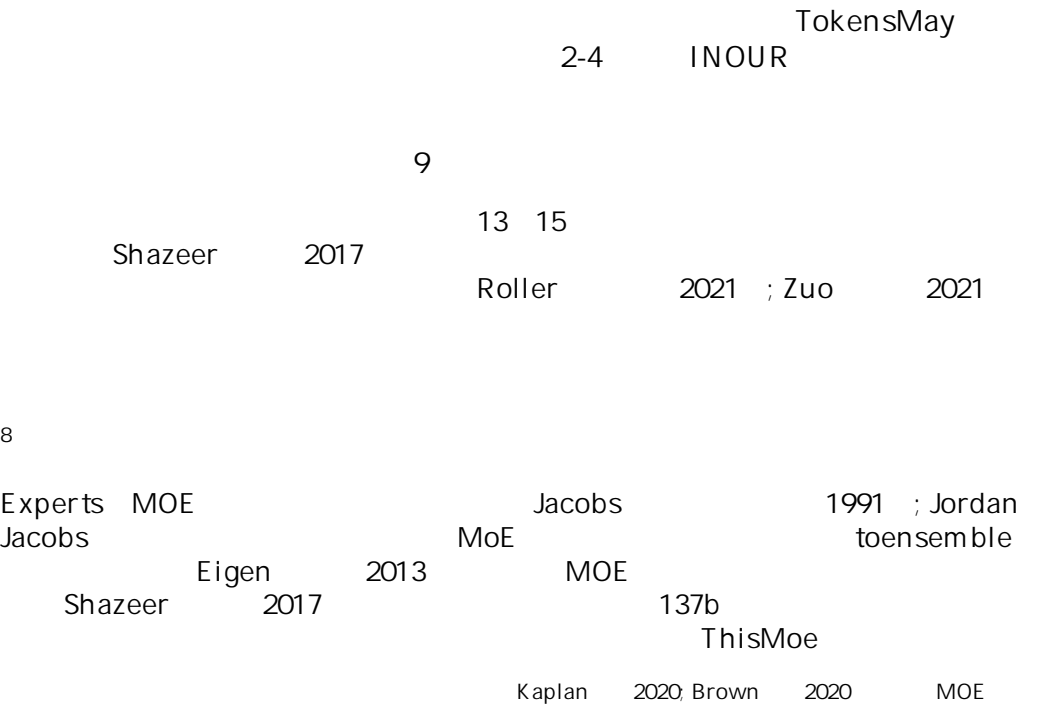
7.1

7.2

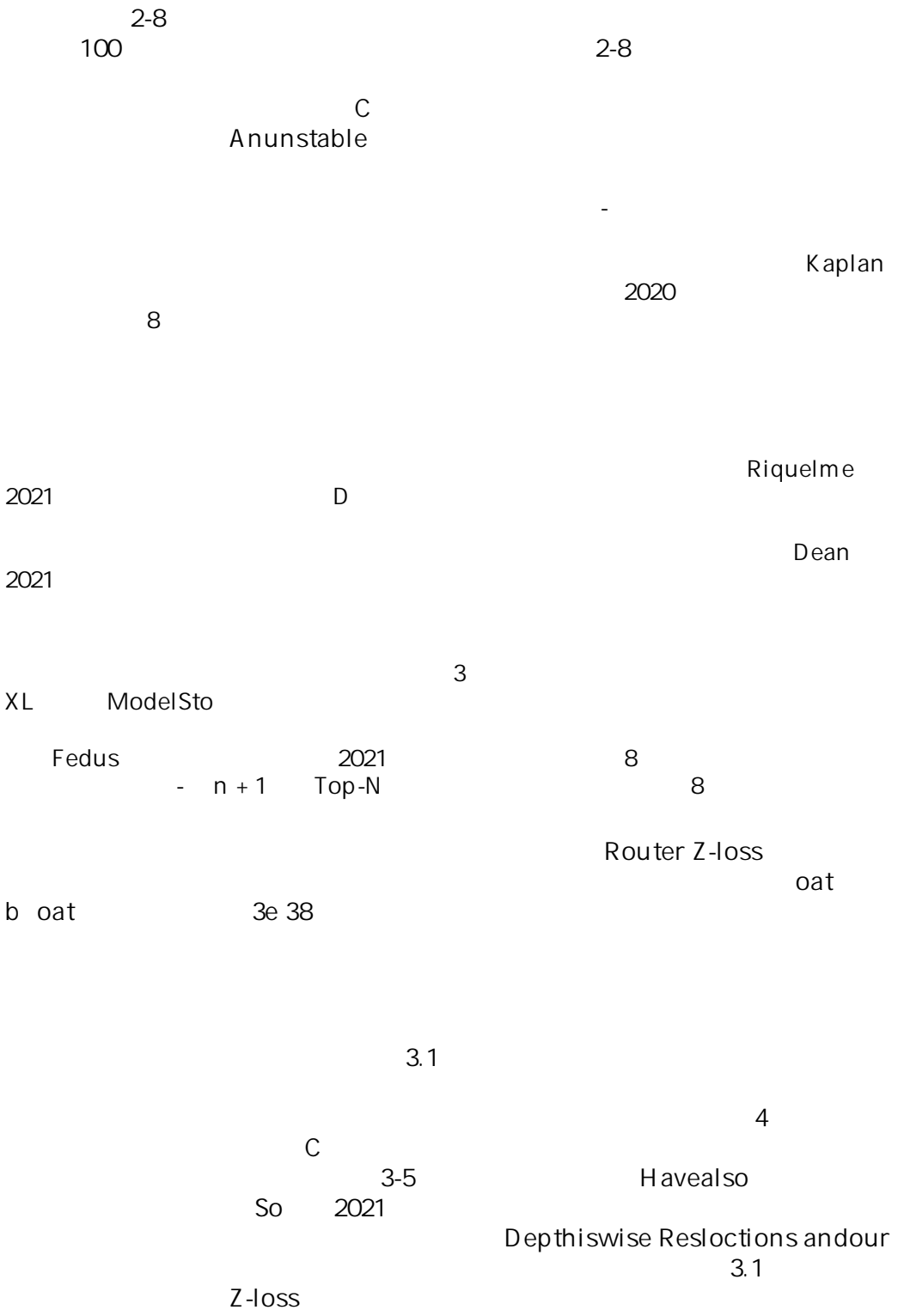
19

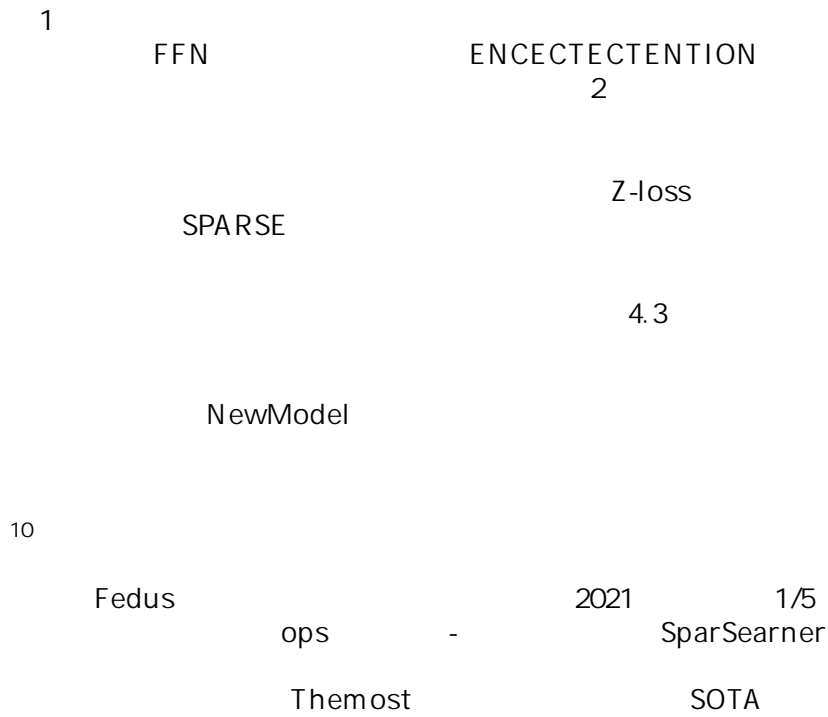
Expert specialization	Routed tokens
Sentinel tokens	to <extra_id.6>to til <extra_id.9> <extra_id.10>to <extra_id.14><extra_id.17> <extra_id.19><extra_id.20><extra_id.21>...
Numbers	\$50 comment .10.2016 ! 20 20 3 ! 5 1. ! 91 ? né ? 2 17 4 17 11 17 8 & 11 & 22:30 02 2016.) iOS
Conjunctions & Articles	of of of their their of any this this your your am von this of Do of of This these our „ „ Ž „ („ („ le les Le la di la sur sur 136 sur n n Y < n h D F n W
Prepositions & Conjunctions	For for or for for or for from because https during https v Ēpar c Pour à a par trè pour pour pour pour c h „ n k g g j n g - and and + c between and and
Proper names	Life Apple iOS A IGT « HB F HB A K A OPP OK HB A Gia C Gia C P Scand Wi G H Z PC G Z Ĩ α PC G Ti CPU PC PC A - Ā Ē OS

Table 15: **Examples of specialization in multilingual experts (encoder).** Multilingual experts also exhibit specialization, which sometimes spans across different languages (e.g. "for" and "pour"). Experts trained on multilingual mixtures do not exhibit language specialization.



Du 2021; Lepikhin 2020; Fedus 2021; Yang 2021; Kim 2021; T -
 2021; Attetxe 2021; Zuo et al 2021 Clark 2022
 XXL 4 Ra el 2019 Du 2021 1
 1 GPT- Brown 2020 /3 thespan
 2021; Yang 2021; Du et al 2021 2021 Fedus
 2021 10T Lin
 Fedus Shazeer 2017 Lepikhin 2020 2020 Kim /
 2021 Fedus 2021 2021 ; Narang 2021 ; Artetxe
 Lewis 2021 Alinear
 -
 2022
 Expert 2021 M -T
 TOP-K Hazimeh K TOP-
 Top-K 2021
 Roller 2021 Zuo 2021
 2021
 2 BLEU Kim
 Fan 2021
 Lepikhin 2020
 +1 BLEU
 Riquelme 2021 15b V-Moeto Deng
 2009 Lou 2021
 MOE You 2021a; B Kumatani Severmoe
 MOE Worderror 2021
 Fedus 2021
 Kudugunta 2021b; Zuo 2021
 NewTask
 Kim 2021
 MOE MOE inmultitask
 Ma 2018
 Gururangan 2021
 /





ACKNOWLEDGEMENTS

Alex Passos Ekin Cubuk Margaret Li Noah Constant Oriol Vinyals Basil
 Mustafa Joan Puigcerver Diego de Las Casas Mike Lewis Ryan Sepassi
 Google Brain

REFERENCES

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giri Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Ves Stoyanov. Efficient large scale language modeling with mixtures of experts, 2021.

arXiv:1607.06450, 2016.

Emmanuel Bengio, Pierre-Luc Bacon, Joelle Pineau, Doina Precup, 2016.

Jonathan Berant, Andrew Chou, Roy Frostig, Percy Liang, freebase from Question-Answer, 1533-1544, 2013.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Aidan Clark, Diego de las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, et al. Unified scaling laws for routed language models. *arXiv preprint arXiv:2202.01169*, 2022.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, arc ai, arXiv, arXiv, 1803.05457, 2018.

Yann N Dauphin, Angela Fan, Michael Auli, David Grangier, 933–941, PMLR, 2017.

Sinn und Bedeutung, 23, 107-124, 2019.

Jeff Dean. Introducing pathways: A next-generation ai architecture. *Google AI Blog*, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, ImageNet, ERARCICAL IMAGE, 2009, IEEE, 248-255, IEEE, 2009.

Tim Dettmers, Mike Lewis, Sam Shleifer, Luke Zettlemoyer, 8, Block-WiseQuantization, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Nan Du, Yanping Huang, Andrew M. Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, Barret Zoph, Liam Fedus, Maarten Bosma, Zongwei Zhou, Tao Wang, Yu Emma Wang, Kellie Webster, Marie Pellat, Kevin Robinson, Kathy Meier-Hellstern, Toju Duke, Lucas Dixon, Kun Zhang, Quoc V Le, Yonghui Wu, Zhifeng Chen, and Claire Cui. Glam: Efficient scaling of language models with mixture-of-experts, 2021.

arXiv:1301.3761, 2013.

Angela Fan Shruti Bhosale Holger Schwenk Zhiyi MA Ahmed El-Kishky Siddharth Goyal
 Man-Deep Baines Onur Celebi Guillaume Wenzek Vishrav Chaudhary
 22 107 1-48 2021

William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: Better text generation via filling in the... *arXiv preprint arXiv:1801.07736*, 2018.

arXiv: , 2021 arXiv

Suchin Gururangan Mike Lewis Ari Holtzman Noah A. Smith Luke Zettlemoyer
 Demixlayers 2021

Hussein Hazimeh Aakanksha Chowdhery Maheswaran Sathiamoorthy Yihua Chen
 Rahul Mazumder Lichan Hong Ed H. Chi Dselect-k
 2021

2015

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Karl Moritz Hermann Tomas Kocisky Edward Grefenstette Lasse
 Espeholt Will Kay Mustafa Suleyman Phil Blunsom Garnett
 AdvanceS 28 1693-1701 Curran
 Asso-Ciates Inc. 2015 URL <https://proceedings.neurips.cc/paper/ /le/afdec/cc/f/cd/fd/fd/f/c/-paper.pdf>

(Sepp Hochreiter) (Jürgen Schmidhuber)
 9(8):1735-1780, 1997

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790-2799. PMLR, 2019.

Sergey Ioffe Christian Szegedy

448-456 PMLR 2015

·A· ·J· ·E·
 3(1):79-87, 1991

·A· em 6(2):181-214, 1994

·S· Triviaqa
 arXiv arXiv: , 2017

·B· arXiv arXiv: ,
 2020

Vladimir Karpukhin Barlas Ouz Sewon Min Patrick Lewis Ledell Wu Sergey Edunov
 Danqichen Wen Tau Yih 2020

Young Jin Kim, Ammar Ahmad Awan, Alexandre Muzio, Andres Felipe Cruz Salinas, Liyang Lu, Amr Hendy, Samyam Rajbhandari, Yuxiong He, and Hany Hassan Awadalla. Scalable and efficient moe training for multitask multilingual models, 2021.

arXiv Sentencepiece
 arXiv: , 2018

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. Beyond distillation: Task-level mixture-of-experts for efficient inference. *arXiv preprint arXiv:2110.03742*, 2021a.

Sneha Kudugunta, Yanping Huang, Ankur Bapna, Maxim Krikun, Dmitry Lepikhin, Minh-Thang Luong, and Orhan Firat. Beyond distillation: Task-level mixture-of-experts for efficient inference. *arXiv preprint arXiv:2110.03742*, 2021b.

Kenichi Kumatani Robert Gmyr Felipe Cruz Salinas Linqun Liu Wei Zuo Devang Patel Ericsun Yu Shi 2021

7:453–466 2019

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668*, 2020.

arXiv: , 2021 arXiv

Mike Lewis Shruti Bhosale Tim Dettmers Naman Goyal Luke Zettlemoyer Arxiv ARXIV 2103.16716 2021

arXiv arXiv: , 2021

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-drop: Regularized dropout for neural networks, 2021.

Junyang Lin, An Yang, Jinze Bai, Chang Zhou, Le Jiang, Xianyan Jia, Ang Wang, Jie Zhang, Yong Li, Wei Lin, Jingren Zhou, and Hongxia Yang. M6-10t: A sharing-delinking paradigm for efficient multi-trillion parameter pretraining, 2021.

Yuxuan Lou Fuzhao Xue Zangwei Zheng Yang -MLP MLP 2021

Jiaqi MA Zhe Zhao Xinyang Yi Jilin Chen Lichan Hong Ed H Chi 24 ACMSIGKDD 1930-1939 2018

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

.E. In IcmI 2010

Sharan Narang, Hyung Won Chung, Yi Tay, William Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2102.11972*, 2021.

(Shashi Narayan) -B- (Shay B Cohen) (Mirella Lapata) arXiv arXiv 1808.08745 2018

Arvind Neelakantan Luke Vilnis Quoc V Le Ilya Sutskever Lukasz Kaiser Karol Kurach James Martens arXiv arXiv 1511.06807 2015

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

Long Ouyang Je Wu Xu Jiang Diogo Almeida Carroll L Wainwright Pamela Mishkin
Chongzhang Sandhini Agarwal Katarina Slama Alex Ray
2022

Razvan Pascanu Tomas Mikolov Yoshua Bengio
1310-1318 PMLR 2013

Quoc Le

arXiv arXiv 2104.10350 2021

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kun-coro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d’Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorraine Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis & insights from training gopher, 2021.

Colin Ra el Noam Shazeer Adam Roberts Katherine Lee Sharan Narang Michael Matena
Yangqi Zhou Wei Li Peter J Liu arXiv
arXiv 1910.10683 2019

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

Carlos Riquelme Joan Puigcerver Basil Mustafa Maxim Neumann Rodolphe Jenatton
André Su-Sano Pinto Daniel Keyzers Neil Houlsby ARXIV
ARXIV 2106.05974 2021

arXiv

arXiv: , 2020

Stephen Roller, Sainbayar Sukhbaatar, Arthur Szlam, and Jason Weston. Hash layers for large sparse models. *arXiv preprint arXiv:2106.04426*, 2021.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. A simple recipe for multilingual grammatical error correction. *arXiv preprint arXiv:2106.03830*, 2021.

Keisuke Sakaguchi Ronan Le Bras Chandra Bhagavatula Yejin Choi Winogrande
Winograd AAAI 34 8732–8740 2020

Tim Salimans Durk P Kingma

29 901–909 2016

arXiv preprint arXiv: , 2019

Noam Shazeer. Glu variants improve transformer, 2020.

Noam Shazeer Mitchell Stern Afactor
4596-4604 PMLR 2018

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarczyk, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Noam Shazeer Youlong Cheng Niki Parmar Dustin Tran Ashish Vaswani Penporn
Koanantakool Peter Hawkins Hyoungho Lee Mingsheng Hong Honglei Young
10414-10423 2018

Sam Shleifer, Jason Weston, and Myle Ott. Normformer: Improved transformer pretraining with extra normalization. *arXiv preprint arXiv:2110.09456*, 2021.

David R So Wojciech Matuszek Hanxiao Liu Zihang Dai Noam Shazeer Quoc V
Le ARXIV ARXIV 2109.08668, 2021

Nitish Srivastava Geoffrey E. Hinton Alex Krizhevsky Ilya Sutskever
Ruslan Salakhutdinov. Dropout 15 1 1929-1958
2014 URL http://www.cs.toronto.edu/~rsalakhu/papers/srivastava_15.pdf

Nassim Nicholas Taleb 3 2012

Yi Tay, Mostafa Dehghani, Jinfeng Rao, William Fedus, Samira Abnar, Hyung Won Chung, Sharan Narang, Dani Yogatama, Ashish Vaswani, and Donald Metzler. Scale efficiently: Insights from pre-training and fine-tuning transformers. *arXiv preprint arXiv:2109.10686*, 2021.

Ashish Vaswani Noam Shazeer Niki Parmar Jakob Uszkoreit Llion Jones Aidan N Gomez
Łukasz Kaiser Illia Polosukhin 5998-6008
2017

Alex Wang Yada Pruksachatkun Nikita Nangia Amanpreet Singh Julian
Michael Felix Hill Omer Levy Samuel Bowman Superglue
3266-3280 2019

Shuohuan Wang Yu Sun Yang Xiang Zhihua Wu Siyu Ding Weibao Gong
Shikun Feng Jun-Yuan Shang Yanbin Zhao Chao Pang Ernie S. 3.0 Titan
ARXIV ARXIV 2112.12731, 2021

ACM 52 4 65-76 2009 Roo FINE

Mike Schuster Quoc V Le Mohammad Norouzi Wolfgang
Macherey Maxim Krikun Klaus Macherey
arXiv arXiv 1609.08144 2016

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

An Yang Junyang Lin Rui Men Chang Zhou Le Jiang Xinyan Jia Ang Wang Jie Zhang
Jia-Ming Wang Yong Li Di Zhang Di Zhang Wei Lin Lin Qu Jingren Zhou Hongxia
Yang M-T 2021

Zhao You Shulin Feng Dan Su Dong Yu SpeechMoe
2021a

Zhao You Shulin Feng Dan Su Dong Yu SpeechMoe Experts IM 2021b

Rico Sennrich arXiv arXiv 1910.07467 2019

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, 2020.

Kevin Duh Benjamin Van Durme
arXiv arXiv 1810.12885 2018

Simiao Zuo, Xiaodong Liu, Jian Jiao, Young Jin Kim, Hany Hassan, Ruofei Zhang, Tuo Zhao, and Jianfeng Gao. Taming sparsely activated transformer with stochastic experts, 2021.

Shazeer

$i = 1 \dots n$

t batch b

2017

Tokens across

F p

n

$$\text{loss} = \frac{1}{N} \sum_{i=1}^N f_i P_i \tag{7}$$

f_i

i

$$f_i = \frac{1}{T} \sum_{x \in B} \mathbb{I}[\text{argmax}_i p(x); ig] \tag{8}$$

P_i

i

$$P_i = \frac{1}{T} \sum_{x \in B} p_i(x) \tag{9}$$

n

7

N

p -vector

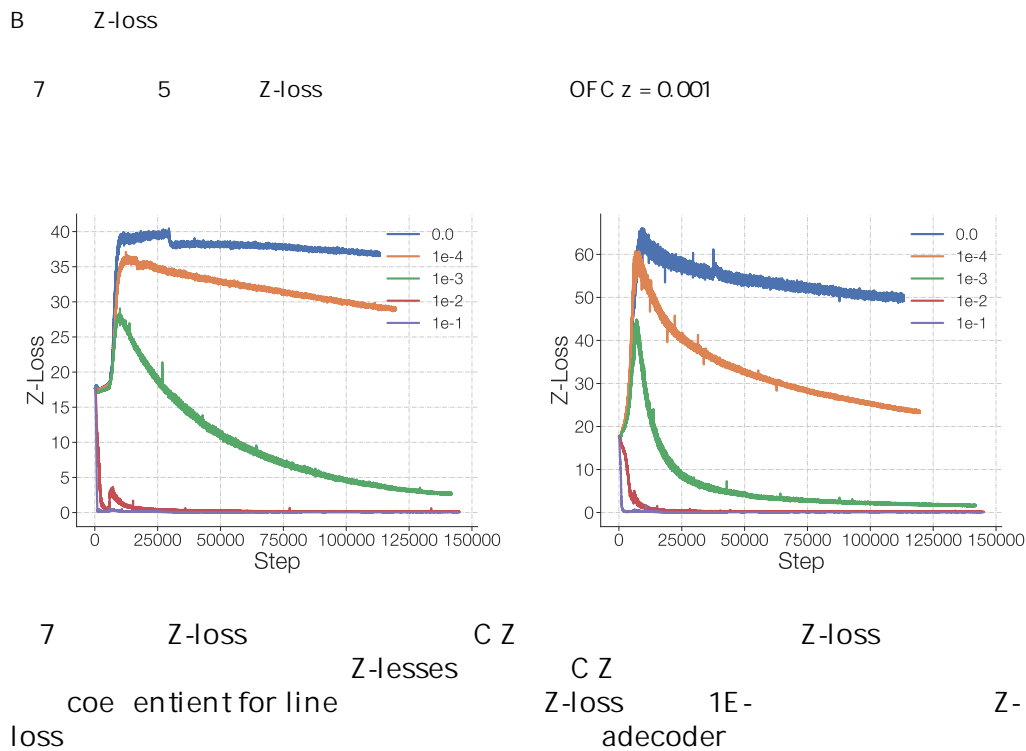
f -vector

n

1/

$\frac{1}{n} - 2$

$= 10$



²A potential source of confusion: $p_i(x)$ is the probability of routing token x to expert i . P_i is the probability fraction to expert i across *all tokens* in the batch B .

c

16	FFN	FFN	FFN	1
				FFN
Model	Neg. Log Perp. (")			
Dense model (baseline)	-1.474	-		
Dense model w/ extra FFN layers	-1.452	0.022		
Sparse model (baseline)	-1.383	-		
Sparse model w/ extra FFN layer <i>after</i> each sparse layer	-1.369	0.014		
Sparse model w/ extra FFN layer <i>before</i> each sparse layer	-1.369	0.014		
Sparse model w/ extra FNN layers placed randomly in the network	-1.376	0.007		

16 FFN FFN 2 FFN

linear n Shazeer 2020 relu n n relu x
 = relu xw 1 w n geglu x = gelu xw 11 2

FFN B [D F F]

tozeros and 11 xW

12 b] w 2

17

Z-loss Weobserve no Multiverative Bias 3.1

4

Model	Neg. Log. Perp. (")	
Dense Baseline	-1.474	-
Sparse Baseline	-1.369	-
Sparse + Additive Bias	-1.371	-0.002
Sparse + Multiplicative Bias	-1.361	0.008

17

Roller

2021

50b+

1B+

J

D

1.0

n

left

Doneto

top- TOP- CF

m space then n > m

ben

Riquelme

2020

BPR BPR

M

18

BPR

2021

the context

Dosovitskiy

TOP-

tokens

M

TOP-

Top-N Bpr routing

BPR

bpr

Algorithm	Train CF	Eval CF	Neg. Log. Perp. (")
Dense	—	—	-1.474
Dense-L	—	—	-1.384
BPR Top-1	0.5	0.5	-1.433
BPR Top-1	0.5	2.0	-1.416
Top-1	0.75	0.75	-1.428
Top-1	0.75	2.0	-1.404
Top-2	0.75	0.75	-1.424
Top-2	0.75	2.0	-1.402
BPR Top-1	0.75	0.75	-1.409
BPR Top-1	0.75	2.0	-1.397
Top-1	1.0	1.0	-1.397
Top-1	1.0	2.0	-1.384
Top-2	1.0	1.0	-1.392
Top-2	1.0	2.0	-1.378
BPR Top-1	1.0	1.0	-1.386
BPR Top-1	1.0	2.0	-1.379
Top-1	1.25	1.25	-1.378
Top-1	1.25	2.0	-1.373
Top-2	1.25	1.25	-1.375
Top-2	1.25	2.0	-1.369
BPR Top-1	1.25	1.25	-1.376
BPR Top-1	1.25	2.0	-1.375

18

TOP-1.25BPR

1BPR TOP-

e

2021

32B

C

Ra el

2019

Glam

Du

Dataset	Tokens (B)	Weight in Mixture
Filtered C4	183	0.17
Filtered Webpages	143	0.34
Wikipedia	3	0.05
Conversations	174	0.23
Forums	247	0.02
Books	390	0.17
News	650	0.02

19

"

Ra el

2019

b

C

Du

2021

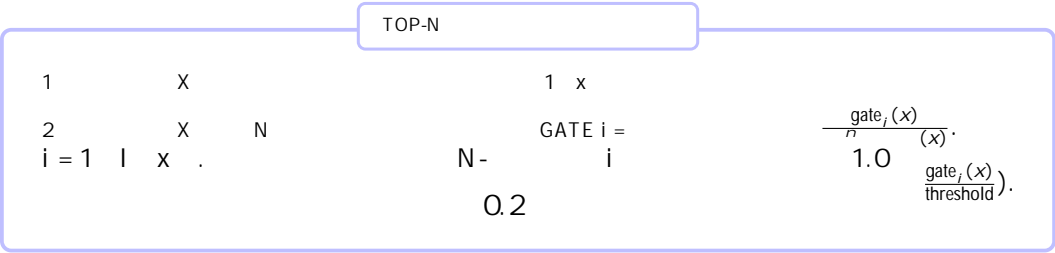
f

20 6 T -LARGE 1M
 op 500kSteps C

Model	Learning Rate	Batch Size	Reset Optimizer Slot Vars	SuperGLUE (%)
Dense	1e-3	1M		84.8
Dense	1e-3	1M	×	84.3
Dense	5e-4	1M		84.8
Dense	5e-4	1M	×	84.2
Dense	1e-4	1M		84.0
Dense	1e-4	1M	×	84.8
Dense	1e-3	262k		84.9
Dense	1e-3	262k	×	83.7
Dense	5e-4	262k		84.9
Dense	5e-4	262k	×	84.0
Dense	1e-4	262k		85.1
Dense	1e-4	262k	×	85.0
Dense	1e-3	65k		83.7
Dense	1e-3	65k	×	82.5
Dense	5e-4	65k		84.4
Dense	5e-4	65k	×	84.1
Dense	1e-4	65k		84.9
Dense	1e-4	65k	×	84.6
Sparse	1e-3	1M		86.9
Sparse	1e-3	1M	×	85.9
Sparse	5e-4	1M		86.1
Sparse	5e-4	1M	×	83.5
Sparse	1e-4	1M		84.3
Sparse	1e-4	1M	×	84.3
Sparse	1e-3	262k		86.2
Sparse	1e-3	262k	×	85.2
Sparse	5e-4	262k		85.5
Sparse	5e-4	262k	×	84.8
Sparse	1e-4	262k		85.1
Sparse	1e-4	262k	×	85.5
Sparse	1e-3	65k		85.8
Sparse	1e-3	65k	×	85.5
Sparse	5e-4	65k		86.5
Sparse	5e-4	65k	×	85.1
Sparse	1e-4	65k		85.6
Sparse	1e-4	65k	×	84.5

20

g



MOE Shazeer 2017; 2018; 2018; Lepikhin et al 2020 MOE TOP- WorksAs works Aws GATE 1 1.0 2 /

0.2 GATE 2

1 2

TOP- N nexperts

0.2 Scorethreshold N-

TOP- TOP- 21 theopposite TOP-

3

0.2 thenext n-

Algorithm	Train CF	Threshold	Neg. Log. Perp. (‘)
Dense	—	—	-1.474
Dense-L	—	—	-1.384
Top-2	3.0	0.2	-1.354
Top-2	3.0	0.05	-1.356
Top-3	3.0	0.2	-1.351
Top-3	3.0	0.05	-1.349

21 TOP- TOP- TOP- TOP-

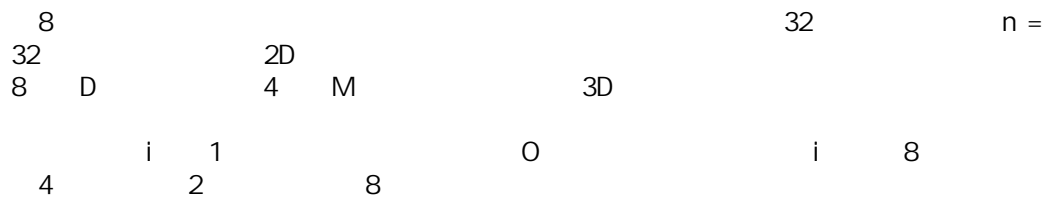
H

Shazeer 2018

d x m 2D d

M n = d x x m

Allreduce each



o x i x m 3D 12 D i d x m 2D mesh

```
Alleduce  All All
          1    AllreduceCalls
```

Parallelism
allreduce
Microbatches

AlIeduce

AlIredue

AL AlIredue All All

DIMension

J

where dropout

Pre-Training

N-
Formoniza-Tion
TOP-N

N
A Doken
N-

Top-N Moe

25 ...

1 5