On the Softmax Bottleneck of Recurrent Language Models : Supplementary Material

Dwarak Govind Parthiban,[1] **Yongyi Mao,** [1] **Diana Inkpen** [1]
[1] University of Ottawa
yottabytt@gmail.com, ymao@uottawa.ca, diana.inkpen@uottawa.ca

1 Tables explicitly referenced in the paper

Model	Test ppl	p-value
Penn Treebank dataset		
Softmax	57.08 ± 0.10	N/A
SS	57.03 ± 0.15	3.92×10^{-1}
GSS	57.02 ± 0.13	2.62×10^{-1}
LMS-PLIF	56.81 ± 0.11	1.91×10^{-5}
MoS	54.88 ± 0.26	2.02×10^{-15}
WikiText-2 dataset		
Softmax	64.63 ± 0.09	N/A
SS	64.35 ± 0.18	3.46×10^{-4}
GSS	64.51 ± 0.13	2.74×10^{-2}
LMS-PLIF	64.15 ± 0.17	2.98×10^{-7}
MoS	61.97 ± 0.43	2.04×10^{-13}

Table 1: p-value resulting from unpaired t-tests between samples of test perplexities of different models and that of the Softmax model. Each model was trained 10 times using 10 randomly sampled seeds for random initialization of the parameters in the model. Values mentioned as $x \pm y$ denote the mean $\pm$ one standard deviation.

Yang et al. (2017)'s cherry-picked contexts					
Context #1	managed properly and with a long-term outlook these can become investment-grade quality properties <eos> canadian <unk> production totaled N metric tons in the week ended oct. N up N N from the preceding week's total of N ____				
Softmax	**tons** **0.90**	million 0.04	metric 0.02	trillion 0.01	billion 0.01
SS	**tons** **0.85**	million 0.04	metric 0.02	billion 0.02	units 0.01
GSS	**tons** **0.68**	units 0.09	million 0.04	billion 0.03	trillion 0.03
LMS-PLIF	**tons** **0.83**	metric 0.11	million 0.02	units 0.01	trillion 0.01
MoS	**tons** **0.40**	million 0.26	billion 0.12	<eos> 0.05	units 0.05
MoC	**tons** **0.36**	million 0.27	billion 0.13	units 0.05	metric 0.04
MoS*	million 0.28	billion 0.23	**tons** **0.19**	trillion 0.10	<eos> 0.05

MoC*	million 0.30	**tons** **0.30**	billion 0.17	<eos> 0.04	trillion 0.03
MoS**	million 0.38	**tons** **0.24**	billion 0.09	barrels 0.06	ounces 0.04
MoC**	billion 0.39	million 0.36	trillion 0.05	<eos> 0.04	N 0.03
Reference #1	canadian <unk> production totaled N metric **tons** in the week ended oct. N up N N from the preceding week 's total of N tons statistics canada a federal agency said <eos>				
Context #2	the thriving <unk> street area offers <unk> of about $ N a square foot as do <unk> locations along lower fifth avenue <eos> by contrast <unk> in the best retail locations in boston san fran- cisco and chicago rarely top $ N ____				
Softmax	million 0.32	billion 0.30	<eos> 0.04	to 0.03	in 0.03
SS	<eos> 0.28	**a** **0.11**	million 0.11	to 0.07	far 0.05
GSS	<eos> 0.18	and 0.13	million 0.10	to 0.08	**a** **0.06**
LMS-PLIF	**a** **0.17**	<eos> 0.14	million 0.07	to 0.07	far 0.06
MoS	million 0.28	billion 0.15	<eos> 0.14	**a** **0.10**	to 0.05
MoC	<eos> 0.12	to 0.11	million 0.10	in 0.08	**a** **0.05**
MoS*	million 0.22	**a** **0.13**	<eos> 0.12	billion 0.11	to 0.07
MoC*	million 0.22	to 0.11	<eos> 0.08	in 0.07	billion 0.06
MoS**	<eos> 0.36	**a** **0.13**	to 0.07	for 0.07	and 0.06
MoC**	million 0.39	billion 0.36	<eos> 0.05	to 0.04	of 0.03
Reference #2	by contrast <unk> in the best retail locations in boston san francisco and chicago rarely top $ N **a** square foot <eos>				
Context #3	as other <unk> governments particularly poland and the soviet union have recently discovered initial steps to open up society can create a momentum for radical change that becomes difficult if not impossible to control <eos> as the days go by the south ____				
Softmax	africa 0.20	**african** **0.11**	to 0.08	korea 0.05	korean 0.05
SS	africa 0.15	**african** **0.14**	korea 0.08	korean 0.05	<unk> 0.05
GSS	africa 0.18	korean 0.10	**african** **0.09**	and 0.06	korea 0.04
LMS-PLIF	africa 0.19	korea 0.05	**african** **0.05**	and 0.05	korean 0.04
MoS	africa 0.15	**african** **0.11**	korea 0.08	of 0.06	and 0.05
MoC	bloc 0.19	africa 0.14	and 0.07	korea 0.06	**african** **0.04**
MoS*	**african** **0.16**	africa 0.13	the 0.08	korea 0.06	<unk> 0.04
MoC*	and 0.11	africa 0.10	bloc 0.07	korea 0.06	<unk> 0.05

MoS**	africa 0.15	**african 0.15**	<eos> 0.14	korea 0.08	korean 0.05
MoC**	<eos> 0.38	and 0.08	of 0.06	or 0.05	<unk> 0.04
Reference #3	as the days go by the south <u>**african**</u> government will be ever more hard pressed to justify the continued <unk> of mr. <unk> as well as the continued banning of the anc and enforcement of the state of emergency <eos>				
Context #4	shares of ual the parent of united airlines were extremely active all day friday reacting to news and rumors about the proposed $ N billion buy-out of the airline by an <unk> group <eos> wall street 's takeover-stock speculators or risk arbitragers had placed unusually large bets that a takeover would succeed and ______				
Softmax	the 0.17	<unk> 0.05	that 0.04	they 0.03	it 0.02
SS	the 0.12	that 0.06	they 0.05	<unk> 0.03	then 0.03
GSS	the 0.17	that 0.04	they 0.04	it 0.03	<unk> 0.03
LMS-PLIF	the 0.10	<unk> 0.05	that 0.03	they 0.02	even 0.02
MoS	the 0.12	<unk> 0.08	**ual 0.08**	that 0.03	coniston 0.02
MoC	the 0.22	<unk> 0.03	they 0.03	a 0.03	then 0.02
MoS*	the 0.10	<unk> 0.07	**ual 0.06**	that 0.03	they 0.02
MoC*	the 0.23	<unk> 0.03	mr . 0.02	**ual 0.03**	that 0.02
MoS**	the 0.14	that 0.07	**ual 0.07**	<unk> 0.03	it 0.02
MoC**	the 0.10	<unk> 0.06	that 0.05	in 0.02	it 0.02
Reference #4	wall street 's takeover-stock speculators or risk arbitragers had placed unusually large bets that a takeover would succeed and <u>**ual**</u> stock would rise <eos>				
Context #5	the government is watching closely to see if their presence in the <unk> leads to increased <unk> protests and violence if it does pretoria will use this as a reason to keep mr. <unk> behind bars <eos> pretoria has n't forgotten why they were all sentenced to life <unk> in the first place for sabotage and ______				
Softmax	<unk> 0.54	political 0.02	violence 0.01	peace 0.01	**conspiracy 0.01**
SS	<unk> 0.45	political 0.02	other 0.01	violence 0.01	incest 0.01
GSS	<unk> 0.45	other 0.01	political 0.01	incest 0.01	civil 0.01
LMS-PLIF	<unk> 0.50	political 0.01	a 0.01	the 0.01	incest 0.01
MoS	<unk> 0.26	violence 0.03	other 0.03	the 0.03	a 0.02
MoC	<unk> 0.65	other 0.02	incest 0.01	acts 0.01	the 0.01

MoS*	<unk> 0.21	acts 0.03	the 0.03	other 0.03	incest 0.02
MoC*	<unk> 0.38	other 0.04	the 0.03	in 0.01	that 0.01
MoS**	<unk> 0.47	violence 0.11	**conspiracy** **0.03**	incest 0.03	civil 0.03
MoC**	<unk> 0.41	the 0.03	a 0.02	other 0.02	in 0.01
Reference #5	pretoria has n't forgotten why they were all sentenced to life <unk> in the first place for sabotage and **<u>conspiracy</u>** to <unk> the government <eos>				
Context #6	china's <unk> <unk> program has achieved some successes in <unk> runaway economic growth and stabilizing prices but has failed to eliminate serious defects in state planning and an <unk> drain on state budgets <eos> the official china daily said retail prices of <unk> foods have n't risen since last december but acknowledged that huge government ____				
Softmax	spending 0.09	costs 0.07	payments 0.04	orders 0.04	sales 0.04
SS	spending 0.10	costs 0.04	<unk> 0.04	orders 0.03	payments 0.03
GSS	spending 0.13	sales 0.07	<unk> 0.04	exports 0.03	officials 0.03
LMS-PLIF	officials 0.10	<unk> 0.05	spending 0.05	and 0.03	contracts 0.03
MoS	spending 0.12	**subsidies** **0.09**	payments 0.08	costs 0.03	sales 0.03
MoC	spending 0.09	debt 0.08	payments 0.08	orders 0.04	<unk> 0.03
MoS*	**subsidies** **0.13**	spending 0.10	benefits 0.04	costs 0.04	orders 0.03
MoC*	spending 0.09	officials 0.05	debt 0.04	**subsidies** **0.03**	<unk> 0.03
MoS**	**subsidies** **0.15**	spending 0.08	officials 0.04	costs 0.04	<unk> 0.04
MoC**	officials 0.04	figures 0.03	efforts 0.03	<unk> 0.03	costs 0.03
Reference #6	the official china daily said retail prices of <unk> foods have n't risen since last december but acknowledged that huge government **<u>subsidies</u>** were a main factor in keeping prices down <eos>				
Our cherry-picked contexts					
Context #1	population drain ends for midwestern states <eos> iowa is making a comeback <eos> so are indiana ohio and michigan <eos> the population of all four ____				
Softmax	<unk> 0.20	of 0.03	companies 0.03	people 0.02	new 0.02
SS	<unk> 0.16	of 0.10	companies 0.08	and 0.02	**states** **0.01**
GSS	companies 0.10	<unk> 0.08	major 0.03	**states** **0.03**	people 0.03
LMS-PLIF	<unk> 0.11	people 0.05	**states** **0.04**	companies 0.04	of 0.03

MoC*	<unk> 0.12	companies 0.05	cities 0.03	small 0.02	major 0.02
MoC	<unk> 0.09	states 0.04	companies 0.04	of 0.04	cities 0.03
MoS*	companies 0.07	<unk> 0.07	**states** **0.06**	areas 0.04	of 0.03
MoS	companies 0.08	<unk> 0.08	areas 0.05	**states** **0.05**	of 0.03
Reference #1	the population of all four **states** is on the <unk> according to new census bureau estimates following declines throughout the early 1980s				
Context #2	the approval of the senate bill was especially sweet for sen. mitchell who had proposed the streamlining <eos> mr. mitchell 's relations with budget director darman who pushed for a capital-gains cut to be added to the measure have been <unk> since mr. darman chose to <unk> the maine democrat and deal with other lawmakers earlier this year during a dispute over drug funding in the fiscal N supplemental spending bill <eos> the deficit reduction ______				
Softmax	is 0.19	would 0.08	was 0.07	in 0.06	has 0.04
SS	is 0.16	was 0.09	in 0.08	would 0.07	has 0.05
GSS	is 0.21	was 0.09	would 0.05	and 0.05	has 0.05
LMS-PLIF	is 0.14	would 0.14	was 0.12	in 0.04	which 0.03
MoC*	is 0.18	was 0.13	would 0.13	will 0.05	has 0.04
MoC	is 0.17	was 0.15	would 0.11	will 0.05	has 0.04
MoS*	is 0.16	was 0.11	would 0.09	in 0.05	has 0.04
MoS	is 0.17	was 0.14	would 0.11	in 0.04	came 0.04
Reference #2	the deficit reduction **bill** contains $ N billion in tax increases in fiscal N and $ N billion over five years				
Context #3	stock prices would still have to go down some additional amount before we become positive on stocks says mr. <unk> president and managing director of renaissance investment management inc. in cincinnati <eos> renaissance which manages about $ N billion drew stiff criticism from many clients earlier this year because it pulled entirely out of ______				
Softmax	the 0.31	its 0.23	a 0.04	<unk> 0.02	their 0.02
SS	the 0.41	its 0.21	a 0.05	<unk> 0.02	it 0.02
GSS	the 0.46	its 0.15	a 0.09	their 0.02	<unk> 0.01
LMS-PLIF	the 0.51	its 0.07	a 0.04	<unk> 0.02	their 0.01
MoC*	the 0.43	its 0.13	a 0.07	<unk> 0.02	program 0.01
MoC	the 0.47	its 0.14	a 0.07	<unk> 0.02	their 0.02
MoS*	the 0.25	its 0.18	a 0.08	<unk> 0.04	an 0.02

MoS	the 0.25	its 0.15	a 0.10	<unk> 0.04	an 0.02
Reference #3	renaissance which manages about $ N billion drew stiff criticism from many clients earlier this year because it pulled entirely out of **stocks** at the beginning of the year and thus missed a strong rally				
Context #4	discount brokerage customers have been in the market somewhat but not whole <unk> like they were two years ago says leslie quick jr. chairman of the quick & <unk> discount brokerage firm <eos> hugo <unk> senior vice president at charles <unk> corp. says schwab ______				
Softmax	's 0.47	& 0.18	is 0.05	has 0.03	was 0.02
SS	& 0.25	's 0.24	is 0.10	has 0.09	was 0.02
GSS	's 0.36	is 0.13	& 0.11	has 0.05	will 0.02
LMS-PLIF	's 0.37	has 0.10	is 0.10	was 0.03	& 0.03
MoC*	's 0.28	& 0.17	is 0.06	has 0.05	<unk> 0.03
MoC	's 0.36	& 0.20	has 0.06	is 0.06	and 0.02
MoS*	& 0.44	's 0.14	has 0.08	is 0.05	was 0.05
MoS	& 0.34	's 0.17	has 0.10	is 0.06	was 0.03
Reference #4	hugo <unk> senior vice president at charles <unk> corp. says schwab **customers** have been neutral to cautious recently about stocks				
Context #5	overall sales of all <unk> vehicles fell N N from a year ago <eos> without gm overall sales for the other u.s. <unk> were roughly flat with N results <eos> some of the u.s. auto makers have already adopted incentives on many N models but they may have to broaden their programs to keep sales up <eos> we 've created a condition where without ______				
Softmax	the 0.13	a 0.12	any 0.10	<unk> 0.07	our 0.02
SS	a 0.15	the 0.13	any 0.08	<unk> 0.07	some 0.02
GSS	a 0.16	the 0.13	any 0.10	<unk> 0.08	an 0.02
LMS-PLIF	a 0.17	the 0.02	any 0.08	<unk> 0.06	some 0.02
MoC*	a 0.18	the 0.17	any 0.06	<unk> 0.04	our 0.03
MoC	a 0.16	the 0.15	any 0.15	<unk> 0.04	our 0.04
MoS*	the 0.13	a 0.12	any 0.07	<unk> 0.05	an 0.02
MoS	a 0.14	the 0.14	any 0.06	<unk> 0.06	our 0.03
Reference #5	we 've created a condition where without **incentives** it 's a tough market said tom kelly sales manager for bill <unk> chevrolet in dearborn mich <eos>				

Context #6	to the extent we lack manpower to staff <unk> jobs in hospitals for example we should raise pay pursue <unk> technology or allow more legal <unk> rather than <unk> high school graduates as short-term workers and cause <unk> among permanent ______				
Softmax	<unk> 0.20	and 0.01	people 0.01	abuse 0.01	care 0.01
SS	<unk> 0.21	and 0.02	groups 0.02	crimes 0.01	<eos> 0.01
GSS	<unk> 0.23	and 0.03	**workers** **0.02**	groups 0.02	standards 0.01
LMS-PLIF	<unk> 0.17	and 0.05	**workers** **0.02**	criteria 0.01	doctors 0.01
MoC*	<unk> 0.31	provisions 0.03	tax 0.02	and 0.02	items 0.02
MoC	<unk> 0.40	crimes 0.02	items 0.02	and 0.01	employment 0.01
MoS*	<unk> 0.10	crimes 0.07	and 0.03	programs 0.02	things 0.02
MoS	<unk> 0.09	crimes 0.05	things 0.04	ones 0.04	criminals 0.03
Reference #6	to the extent we lack manpower to staff <unk> jobs in hospitals for example we should raise pay pursue <unk> technology or allow more legal <unk> rather than <unk> high school graduates as short-term workers and cause <unk> among permanent **workers** paid lesser amounts to do the same jobs				
	Randomly selected contexts				
Context #1	amid a crowd of <unk> stocks <unk> technology inc. 's stock fell particularly hard friday dropping N N because its problems were compounded by disclosure of an unexpected loss for its fiscal first quarter <eos> the <unk> software company said it expects a $ N million net loss for the fiscal first quarter ended sept. N <eos> it said analysts had been expecting a small profit for the period <eos> revenue is ______				
Softmax	**expected** **0.15**	$ 0.15	about 0.07	the 0.04	n't 0.04
SS	**expected** **0.39**	n't 0.04	estimated 0.03	<unk> 0.02	likely 0.02
GSS	**expected** **0.29**	n't 0.07	$ 0.05	estimated 0.05	likely 0.04
LMS-PLIF	**expected** **0.34**	$ 0.08	estimated 0.05	n't 0.04	the 0.03
MoC*	**expected** **0.22**	$ 0.07	n't 0.07	up 0.03	likely 0.02
MoC	**expected** **0.39**	n't 0.05	$ 0.03	estimated 0.03	likely 0.02
MoS*	**expected** **0.16**	$ 0.15	n't 0.06	the 0.04	flat 0.04
MoS	**expected** **0.21**	$ 0.08	n't 0.06	likely 0.04	the 0.03
Reference #1	revenue is **expected** to be up modestly from the $ N million reported a year ago				
Context #2	centrust however <unk> the branch sale saying it would bring in $ N million and reduce the thrift 's assets to $ N billion from $ N billion <eos> it said the sale would give it positive tangible capital of $ N million or about N N of assets from a negative $ N million as of sept. N thus bringing ______				

Softmax	the 0.39	its 0.08	a 0.07	it 0.05	$ 0.04
SS	the 0.55	a 0.04	it 0.04	its 0.03	<unk> 0.02
GSS	the 0.48	its 0.06	a 0.04	it 0.04	to 0.03
LMS-PLIF	the 0.50	its 0.06	a 0.04	it 0.04	<unk> 0.02
MoC*	the 0.28	$ 0.08	a 0.06	it 0.03	its 0.03
MoC	the 0.39	to 0.08	$ 0.05	its 0.05	a 0.03
MoS*	the 0.39	a 0.10	its 0.05	it 0.03	an 0.02
MoS	the 0.34	a 0.12	$ 0.05	its 0.05	more 0.03
Reference #2	it said the sale would give it positive tangible capital of $ N million or about N N of assets from a negative $ N million as of sept. N thus bringing **centrust** close to regulatory standards <eos>				
Context #3	national also <unk> in the northwest <unk> program along with four other airlines including delta and usair group inc. 's usair unit <eos> a month ago hertz of park <unk> n.j. said that it would drop its marketing agreements at year end with delta america west and texas air corp. 'S continental airlines and eastern airlines and that <unk> with american airlines ual inc 's united airlines and ______				
Softmax	the 0.15	<unk> 0.08	united 0.05	american 0.03	eastern 0.03
SS	the 0.19	its 0.11	<unk> 0.05	american 0.04	amr 0.04
GSS	amr 0.10	the 0.09	ual 0.06	united 0.03	<unk> 0.03
LMS-PLIF	the 0.18	ual 0.05	<unk> 0.05	amr 0.04	its 0.03
MoC*	the 0.13	united 0.08	its 0.06	<unk> 0.06	amr 0.05
MoC	united 0.10	american 0.10	the 0.09	amr 0.09	trans 0.06
MoS*	the 0.13	<unk> 0.09	amr 0.06	its 0.05	trans 0.04
MoS	the 0.14	<unk> 0.09	trans 0.08	amr 0.04	texas 0.04
Reference #3	a month ago hertz of park <unk> n.j. said that it would drop its marketing agreements at year end with delta america west and texas air corp. 'S continental airlines and eastern airlines and that <unk> with american airlines ual inc 's united airlines and **usair** also would be ended sometime after dec. N				
Context #4	in a filing with the securities and exchange commission mr. <unk> <unk> said <unk> syndicate inc. <unk> ii inc. and <unk> iii inc. bought the N shares on oct. N for $ N million or $ N a share <eos> mr. <unk> <unk> said that he <unk> group ltd. <unk> <unk> ii and <unk> iii are all affiliated and hold a combined ______				
Softmax	N 0.20	**stake** **0.01**	shares 0.01	$ 0.01	<unk> 0.01

SS	N 0.20	**stake** **0.01**	number 0.01	interest 0.01	equity 0.01
GSS	N 0.20	**stake** **0.01**	price 0.01	<unk> 0.01	number 0.01
LMS-PLIF	N 0.20	**stake** **0.01**	share 0.01	$ 0.01	<unk> 0.01
MoC*	N 0.20	**stake** **0.01**	number 0.01	$ 0.01	profit 0.01
MoC	N 0.20	**stake** **0.01**	$ 0.01	number 0.01	share 0.01
MoS*	N 0.20	**stake** **0.01**	number 0.01	<unk> 0.01	$ 0.01
MoS	N 0.20	**stake** **0.01**	<unk> 0.01	$ 0.01	company 0.01
Reference #4	mr. <unk> <unk> said that he <unk> group ltd. <unk> <unk> ii and <unk> iii are all affiliated and hold a combined **stake** of N shares or N N				
Context #5	the key u.s. and foreign annual interest rates below are a guide to general levels but do n't always represent actual transactions <eos> prime rate <eos> N N N <eos> the base ______				
Softmax	**rate** **0.20**	rates 0.01	base 0.01	unit 0.01	on 0.01
SS	**rate** **0.20**	on 0.01	rates 0.01	yield 0.01	of 0.01
GSS	**rate** **0.20**	rates 0.01	on 0.01	of 0.01	yield 0.01
LMS-PLIF	**rate** **0.20**	on 0.01	rates 0.01	charge 0.01	base 0.01
MoC*	**rate** **0.20**	on 0.01	rates 0.01	yield 0.01	of 0.01
MoC	**rate** **0.20**	on 0.01	rates 0.01	of 0.01	yield 0.01
MoS*	**rate** **0.20**	on 0.01	of 0.01	<unk> 0.01	base 0.01
MoS	**rate** **0.20**	of 0.01	<unk> 0.01	base 0.01	lending 0.01
Reference #5	the base **rate** on corporate loans at large u.s. money center commercial banks				
Context #6	enthusiasts assume that national service would get important work done <unk> forest fires fought housing <unk> students <unk> <unk> centers <unk> <eos> there is important work to be done and existing service and conservation corps have shown that even <unk> who start with few ______				
Softmax	<unk> 0.20	of 0.01	other 0.01	new 0.01	groups 0.01
SS	<unk> 0.20	other 0.01	of 0.01	people 0.01	new 0.01
GSS	<unk> 0.20	of 0.01	other 0.01	new 0.01	people 0.01
LMS-PLIF	<unk> 0.20	of 0.01	other 0.01	new 0.01	people 0.01
MoC*	<unk> 0.20	of 0.01	people 0.01	other 0.01	new 0.01

MoC	<unk> 0.20	of 0.01	other 0.01	people 0.01	others 0.01
MoS*	<unk> 0.20	of 0.01	people 0.01	other 0.01	new 0.01
MoS	<unk> 0.20	of 0.01	people 0.01	other 0.01	new 0.01
Reference #6	there is important work to be done and existing service and conservation corps have shown that even <unk> who start with few **skills** can do much of it well but not <unk>				

Table 2: Qualitative analysis (Top-5 predictions made by each model for next-token conditioned on a context). MoS** and MoC** are the results reported in (Yang et al. 2017) that use NT-ASGD. Our reproduced versions MoS* and MoC* also use NT-ASGD.

Word similarity benchmark	**Softmax**	**SS**	**GSS**	**LMS-PLIF**	**MoS**	**MoC**
Learned embeddings from language models trained on PTB						
WS-353	0.4160	0.3968	0.3949	0.4167	0.3609	0.4025
WS-353-SIM	0.4550	0.4462	0.4507	0.4710	0.3846	0.4451
WS-353-REL	0.3774	0.3491	0.3470	0.3714	0.3399	0.3361
RG-65	0.3697	0.5030	0.5152	0.5152	0.2485	0.6121
MC-30	0.3833	0.4667	0.3833	0.3500	0.1333	0.4167
MTurk-287	0.6086	0.6153	0.5918	0.5843	0.6171	0.5857
MTurk-771	0.4273	0.4341	0.4378	0.4199	0.3985	0.4186
MEN	0.4299	0.4460	0.4355	0.4298	0.3789	0.4337
YP-130	0.1734	0.1657	0.1190	0.1279	0.2817	0.2780
VERB-143	0.4388	0.4350	0.4599	0.4534	0.4672	0.4358
RW-STANFORD	0.4787	0.4676	0.4527	0.4819	0.4904	0.4603
SimVerb-3500	0.1185	0.1212	0.1260	0.1161	0.1133	0.1331
SimLex-999	0.2273	0.2067	0.2361	0.2060	0.1887	0.1950
Learned embeddings from language models trained on WT2						
WS-353	0.4658	0.4691	0.4799	0.4657	0.4155	0.4676
WS-353-SIM	0.5925	0.6007	0.6077	0.6022	0.5551	0.5872
WS-353-REL	0.3759	0.3905	0.3933	0.3654	0.3238	0.3777
RG-65	0.5701	0.5368	0.5547	0.5231	0.4868	0.5426
MC-30	0.7308	0.7627	0.7442	0.7247	0.6050	0.7490
MTurk-287	0.5405	0.5682	0.5634	0.5485	0.5685	0.5068
MTurk-771	0.4483	0.4559	0.4581	0.4450	0.4129	0.4425
MEN	0.5895	0.5883	0.5965	0.5830	0.5399	0.5659
YP-130	0.1889	0.2127	0.2388	0.2272	0.1665	0.2117
VERB-143	0.4268	0.4306	0.4401	0.4253	0.4541	0.4646
RW-STANFORD	0.4565	0.4698	0.4582	0.4521	0.4487	0.4781
SimVerb-3500	0.1243	0.1283	0.1288	0.1283	0.1438	0.1515
SimLex-999	0.2432	0.2337	0.2276	0.2325	0.1783	0.2175

Table 3: Spearman's rank correlation coefficient ρ values on different word similarity benchmarks for learned word embeddings from language models trained on PTB and WT2 datasets

K	#Param	Train ppl	Test ppl	Rank
1	19.05M	56.42	64.50	282
3	19.40M	41.77	59.25	5,575
5	19.75M	38.09	58.38	8,057
10	20.62M	35.48	56.21	9,976
15	21.50M	33.08	56.07	9,979
20	22.37M	32.19	56.19	9,980

Table 4: MoS model on PTB dataset for different number of mixtures K.

K	#Param	Train ppl	Test ppl	Rank
10	33.92M	36.94	63.62	12,198
15	34.90M	35.92	63.06	13,229
20	35.88M	35.35	62.76	13,998

Table 5: MoS model on WT2 dataset for different number of mixtures K.

2 Details about hyperparameters and hyperparameter finetuning

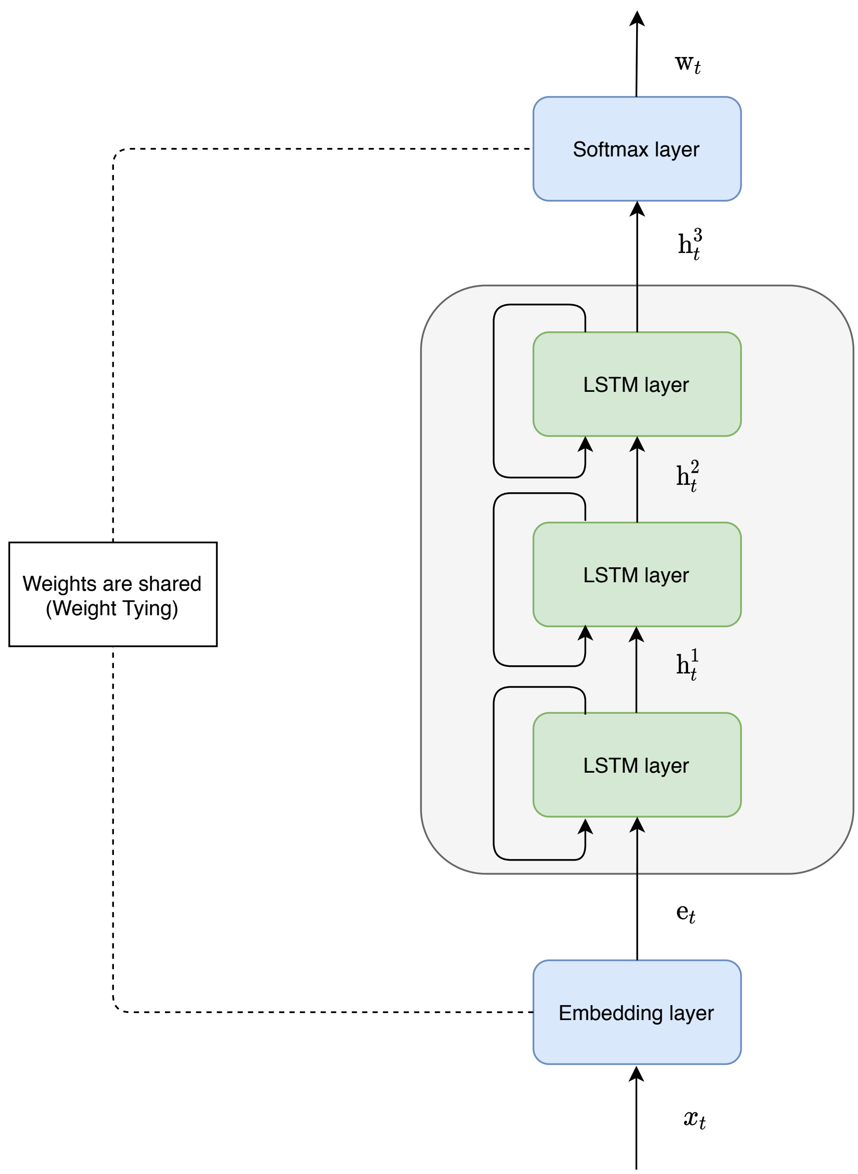

Figure 1: A rolled-up AWD-LSTM network

2.1 Notable differences in hyperparameters for models under comparison

We used the hyperparameters reported by Merity, Keskar, and Socher (2018) for Softmax, SS, GSS, and LMS-PLIF models. For MoS and MoC, we used the hyperparamaters reported by Yang et al. (2017). Some of the notable differences among these two sets of hyperparameters are shown in Table 6. The common hyperparameter values in both these sets are 0.1 for embedding matrix dropout, 0.4 for dropout on $\mathbf{h}_t^3$ (Figure 1), 0.5 for dropconnect on LSTM weights, 1.2e-6 for scaling factor of L2 regularization (weight decay), 2 and 1 for scaling factors of activation and temporal activation regularization.

Hyperparameter	For Softmax, SS, GSS, and LMS-PLIF		For MoS and MoC	
	Data set		**Data set**	
	PTB	WT2	PTB	WT2
Dropout for $\mathbf{e}_t$	0.4	0.65	0.4	0.55
Dropout for $\mathbf{h}_t^1, \mathbf{h}_t^2$	0.3	0.2	0.225	0.2
Learning rate	20.0	30.0	20.0	15.0
Batch size	20	80	12	15
Random seed	141	1,881	28	1,881
$dim(\mathbf{e}_t)$	400	400	280	300
$dim(\mathbf{h}_t^1), dim(\mathbf{h}_t^2)$	1,150	1,150	960	1,150
$dim(\mathbf{h}_t^3)$	400	400	620	650

Table 6: Differences in hyperparameters. $dim(.)$ denotes the dimension. Refer Figure 1 to know about $\mathbf{e}_t$, $\mathbf{h}_t^1$, $\mathbf{h}_t^2$, $\mathbf{h}_t^3$

2.2 Hyperparameters specific to the PLIF layer and the MoS layer

As LMS-PLIF and MoS models has extra trainable parameters in the form of PLIF and MoS layers, there are a few hyperparameters which are exclusive to them. We obtained the values of the hyperparameters exclusive to the PLIF layer through a discussion with Ganea et al. (2019). For the PLIF layer (Ganea et al. 2019), on both datasets, K was set to 10^5; T was set to 20; a layer specific learning rate of 0.02 was used. For the MoS layer, as reported by Yang et al. (2017), on both datasets, a dropout of 0.29 was used and the number of mixtures $K = 15$ was used.

2.3 Hyperparameter finetuning for comparing MoS and Softmax models

From the works of (Wang, Gong, and Liu 2019; Wang et al. 2020), we came to know that adding small gaussian noise to $\mathbf{e}_t$ (Figure 1) helps in better performance. Hence, we also included this as a hyperparameter (which we call embedding noise) in the set of hyperparameters that we used for making the Softmax model to perform as good as that of the MoS model on PTB dataset. As shown in Tables 7 and 8, we finetuned only a total of six hyperparameters in two stages, and used the best performing hyperparameter values for MoS† and Softmax‡ models. The cross product of the set of values for hyperparameters were used for the search.

Hyperparameter	**Values used**
Dropout for $\mathbf{e}_t$	0.2, 0.4
Dropout for $\mathbf{h}_t^1, \mathbf{h}_t^2$	0.225, 0.3
Embedding noise	0.10, 0.15

Table 7: First stage of hyperparameter finetuning for both MoS and Softmax models

Hyperparameter	Values used
Embedding matrix dropout	[0.075, 0.125](0.025)
Dropout for $\mathbf{e}_t$	[0.28, 0.34](0.01)
Dropout for $\mathbf{h}_t^1, \mathbf{h}_t^2$	[0.20, 0.35](0.025)
Dropout for $\mathbf{h}_t^3$	0.26
Embedding noise	[0.10, 0.20](0.025)
Weight decay	[1.2e-6,1.5e-6](0.1e-6)

Table 8: Second stage of hyperparameter finetuning for both MoS and Softmax models. [x,y](z) denote the values between x and y with a step size of z.

Hyperparameter	Final value	
	Softmax‡	MoS†
Embedding matrix dropout	0.125	0.1
Dropout for $\mathbf{e}_t$	0.28	0.4
Dropout for $\mathbf{h}_t^1, \mathbf{h}_t^2$	0.225	0.225
Dropout for $\mathbf{h}_t^3$	0.26	0.4
Embedding noise	0.15	0.10
Weight decay	1.5e-6	1.2e-6

Table 9: Best performing hyperparameter values after two stages of finetuning for MoS† and Softmax‡ models.

3 Other supporting tables for claims made in the paper

The performance differences when ET-ASGD (epoch number 200) is used over NT-ASGD (non monotone interval 5) for models on both PTB and WT2 datasets are shown in Tables 10 and 11 respectively.

Model	#Param	Train ppl	Validation ppl	Test ppl	Rank
NT-ASGD					
Softmax	24.22M	34.05	60.35	58.07	402
SS	24.22M	33.68	60.45	57.75	4,906
GSS	24.22M	34.24	59.95	57.60	8,276
LMS-PLIF	24.32M	37.19	60.86	58.45	510
MoS	21.50M	33.08	58.21	56.07	9,979
MoC	21.50M	33.73	59.84	57.40	282
ET-ASGD					
Softmax	24.22M	34.03	59.48	57.10	402
SS	24.22M	32.83	59.95	57.16	4,979
GSS	24.22M	34.21	59.37	56.78	8,989
LMS-PLIF	24.32M	37.07	59.08	56.67	580
MoS	21.50M	31.62	57.12	55.11	9,983
MoC	21.50M	31.37	58.38	55.81	282

Table 10: Performance comparison for NT-ASGD vs ET-ASGD on PTB

Model	#Param	Train ppl	Validation ppl	Test ppl	Rank
NT-ASGD					
Softmax	33.55M	39.07	68.35	65.28	402
SS	33.55M	39.21	67.84	65.08	5,879
GSS	33.55M	39.05	67.72	65.07	9,130
LMS-PLIF	33.65M	41.11	68.54	65.59	479
MoS	34.90M	35.92	65.93	63.06	13,215
MoC	34.90M	37.21	69.08	66.42	302
ET-ASGD					
Softmax	33.55M	39.09	67.59	64.56	402
SS	33.55M	39.19	67.19	64.33	6,590
GSS	33.55M	39.12	66.97	64.38	10,145
LMS-PLIF	33.65M	41.19	67.19	64.32	513
MoS	34.90M	35.99	64.58	61.90	15,738
MoC	34.90M	37.23	68.19	65.83	302

Table 11: Performance comparison for NT-ASGD vs ET-ASGD on WT2

We showed, for MoS models, that rank can be increased without increasing the number of mixtures but by adjusting the dropout rates of the MoS layer. The complete results for that experiment on both PTB and WT2 datasets are shown in Table 12.

Dropout	Train ppl	Test ppl	Rank
Penn Treebank dataset			
0.29	33.08	56.07	9,979
0.145	29.21	59.09	9,985
0.00	23.81	64.82	9,992
WikiText-2 dataset			
0.29	39.11	63.06	13,215
0.145	32.19	64.38	17,256
0.00	27.51	68.49	19,427

Table 12: MoS model for different dropout rates applied to the MoS layer. All the models use 15 mixtures.

4 Other relevant observations

4.1 About word similarity benchmarks

As the vocabulary sizes of PTB and WT2 datasets are $10,000$ and $33,278$ respectively, it can be understood that not all word pairs in the benchmarks can be present in the vocabulary. A brief summary about this statistics is shown in Table 13.

Dataset	# Pairs	# Pairs not in PTB	# Pairs not in WT2
WS-353	353	116	48
WS-353-SIM	203	66	28
WS-353-REL	252	80	28
RG-65	65	55	20
MC-30	30	21	4
MTurk-287	287	146	106
MTurk-771	771	346	99
MEN	3,000	1,952	863
YP-130	130	65	43
VERB-143	144	9	0
RW-STANFORD	2,034	1,889	1,605
SimVerb-3500	3,500	1,746	1,080
SimLex-999	999	424	106

Table 13: Word pairs in benchmarks vs those in the vocabularies of PTB and WT2 datasets

4.2 Log scale vs normalized linear scale

Note that $\log P$ matrix is denoted as $\mathbf{Q}_\theta$.

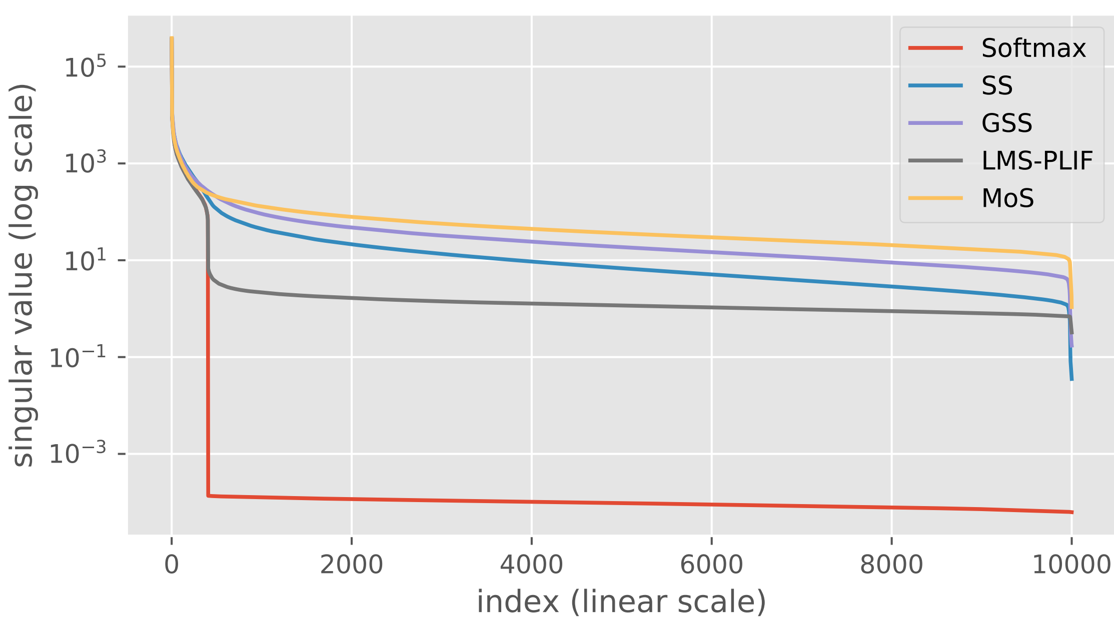

Figure 2: Singular values of $\mathbf{Q}_\theta$ on PTB's test set.

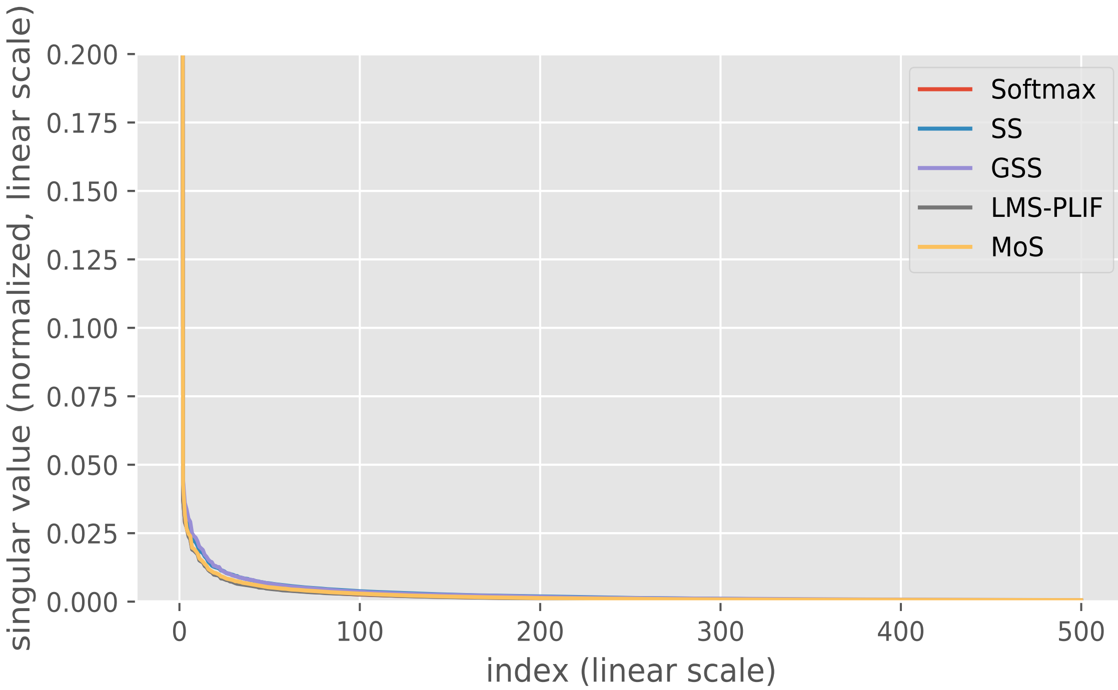

Figure 3: Normalized singular values [0,1] of $\mathbf{Q}_\theta$ on PTB's test set. For better visibility, x-axis limited to show first 500 indices and y-axis limited to show [0, 0.2].

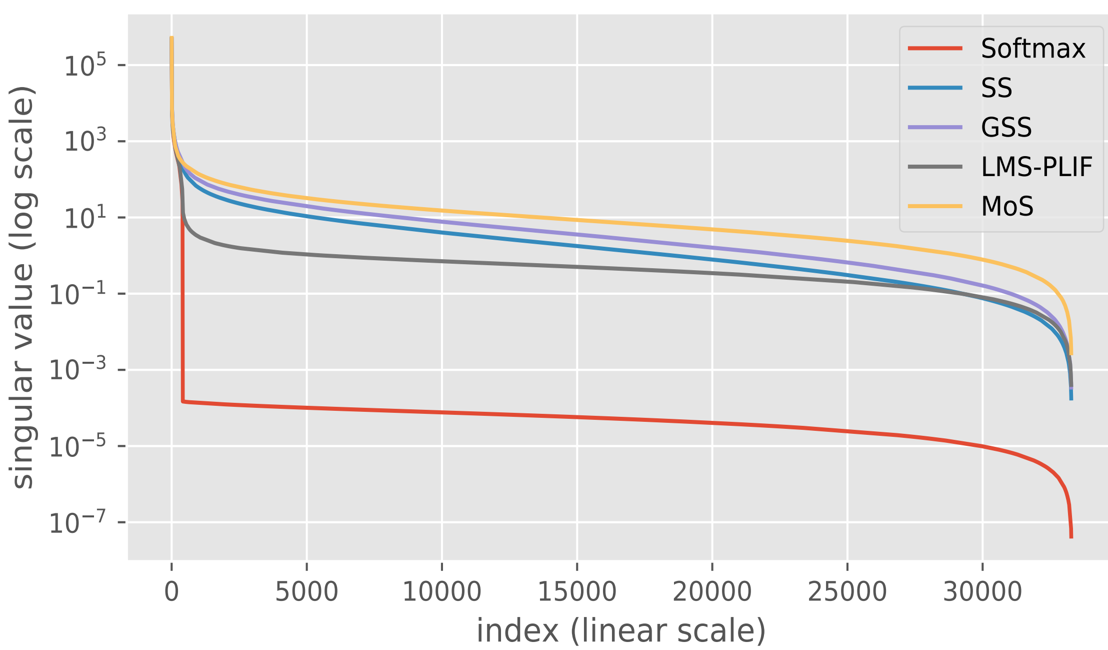

Figure 4: Singular values of $\mathbf{Q}_\theta$ on WT2's test set.

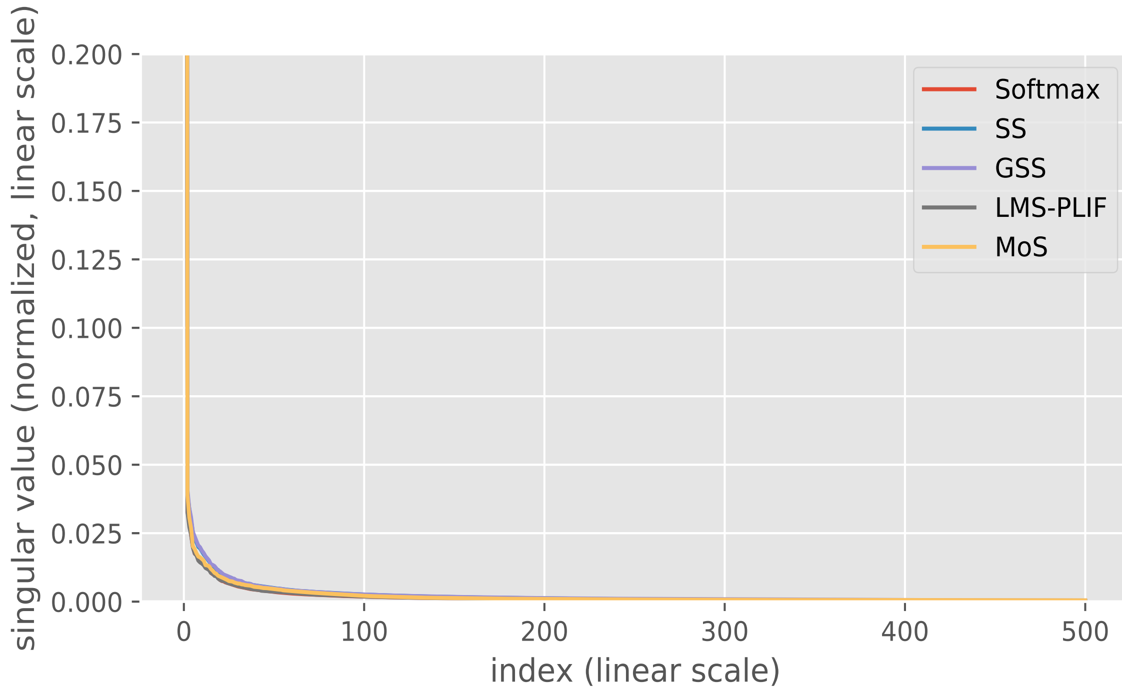

Figure 5: Normalized singular values [0,1] of $\mathbf{Q}_\theta$ on WT2's test set. For better visibility, x-axis limited to show first 500 indices and y-axis limited to show [0, 0.2].

5 Press' rank vs ϵ-effective rank

Note that $\log P$ matrix is denoted as $\mathbf{Q}_\theta$.

5.1 Robustness of Press' Rank

As the rank values for large real-valued matrices are typically calculated using Press et al. (2007)'s approach (which we call as Press' rank), we wanted to check how robust is it to tiny noises. We conducted two experiments to check the sensitivity of Press' rank. For both experiments we use MoS's $\mathbf{Q}_\theta$ constructed from PTB's test set. We know that $\mathbf{Q}_\theta \in \mathbb{R}^{82,430\times 10,000}$ as there are 82,430 contexts in the test set and 10,000 words in the vocabulary of PTB. We need both $\mathbf{U}$ and $\mathbf{V}$ in addition to $\mathbf{S}$ as outputs from SVD routine for our experiments. However, in practice, computing $\mathbf{U}$ and $\mathbf{V}$ in addition to $\mathbf{S}$ from Thin SVD routines is very time consuming when compared to only computing $\mathbf{S}$. Hence, we consider only the top 10,000 contexts instead of 82,430 instances for these experiments, i.e., $\mathbf{Q}_\theta \in \mathbb{R}^{10,000\times 10,000}$. For this truncated $\mathbf{Q}_\theta$, the rank is 8,933 when calculated using Press et al. (2007)'s approach. Remember that the rank of $\mathbf{Q}_\theta$ for MoS was 9,979 (as seen in Table 10) when all the contexts from the test set were used to construct $\mathbf{Q}_\theta$. But this difference in rank does not impact our experiments as we just need a $\mathbf{Q}_\theta$ that occurs in practice rather than a randomly generated matrix.

5.2 To Reconstruction Errors

Algorithm 1

Require:
$\mathbf{Q}_\theta \in \mathbb{R}^{m \times n}, m \geq n$

$\mathbf{U}, \mathbf{S}, \mathbf{V}^T \leftarrow svd(\mathbf{Q}_\theta)$
$\mathbf{S}_p \leftarrow prune(\mathbf{S}, 10)$ ▷ retains only top-10 values. puts the rest to 0.
$i \leftarrow 10$
while $i \neq 0$ **do**
  $\mathbf{U}_q, \mathbf{S}_q, \mathbf{V}_q^T \leftarrow svd(\mathbf{S}_p)$
  $\mathbf{S}_p = \mathbf{U}_q \mathbf{S}_q \mathbf{V}_q^T$
  $print(rank(\mathbf{S}_p))$ ▷ $rank(.)$ uses Press et. al (Press et al. 2007)'s approach.
  $i \leftarrow i - 1$
end while

As we want the noise to be something that is unavoidable and that occurs naturally, we check for the robustness of Press' rank to errors that arise when reconstructing the original matrix from SVD factorized matrices. First, we do an SVD for $\mathbf{Q}_\theta$ to calculate the matrices $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{V}^T$. We keep the top-10 singular values of $\mathbf{S}$ and set the rest to zero. Now, the product $\mathbf{USV}^T$ matrix is of rank 10. We repeatedly factorize and reconstruct this low-rank matrix and check if its rank increases when calculated using Press et al. (2007)'s approach. The steps of the algorithm are outlined in Algorithm 1. The results for this experiment are shown in Table 14. Though the reconstruction error increases after every iteration, Press' rank seems to be robust.

Iteration	1	2	3	4	5	6	7	8	9	10
Rank	10	10	10	10	10	10	10	10	10	10
Error $\times 10^{-28}$	0.97	0.99	1.01	1.05	1.08	1.10	1.13	1.17	1.23	1.28

Table 14: Robustness to SVD reconstruction errors.

5.3 To Gaussian Additive Noise

Algorithm 2

Require:
$\mathbf{Q}_\theta \in \mathbb{R}^{m \times n}, m \geq n$
$c, s \in \mathbb{R}$
$j \in \mathbb{Z}$

$\mathbf{U}, \mathbf{S}, \mathbf{V}^T \leftarrow svd(\mathbf{Q}_\theta)$ ▷ $\mathbf{S} \in \mathbb{R}^{n \times n}$
$\mathbf{S}_p \leftarrow prune(\mathbf{S}, 10)$ ▷ retains only top-10 values. makes the rest as 0.
while $j \neq 0$ **do**
  $\mathbf{Z} \leftarrow gaussian\_random(0, c/n^2)$
  $\mathbf{S}_g = \mathbf{S}_p + \mathbf{Z}$
  $print(rank(\mathbf{S}_g))$ ▷ $rank(.)$ uses Press et. al (Press et al. 2007)'s approach.
  $c \leftarrow c + s$
  $j \leftarrow j - 1$
end while

Now, we check for the robustness of Press' rank to very small gaussian additive noises. The steps of the algorithm are outlined in Algorithm 2. The procedure is same as our previous experiment till the step of making the rank of the product matrix $\mathbf{USV}^T$ equal to 10. The parameter c which is fed as an input to the algorithm controls the standard deviation of the zero-mean gaussian distribution from which the additive noise is sampled. Iteratively, c is changed according to the step size s which is another parameter of the algorithm. For different additive noises, we look for change in Press' rank for the matrix which has an actual rank of 10.

We inspect the behavior of Press' rank by using different step sizes. The results for different c's are shown in Table 15. Results for step sizes in the power of 2 are shown in the first sub-table, for step size of 2 in the second sub-table, and for step size of 1 in the third sub-table. We can see that for c's lower than 10^{-25}, the rank of the matrix is computed correctly using Press et al. (2007)'s approach. However, for just a few tens of higher magnitudes, the value of Press' rank becomes close to full-rank.

c	10^{-2}	10^{-4}	10^{-8}	10^{-16}	10^{-32}
Rank	10,000	10,000	10,000	9,998	10

c	10^{-20}	10^{-22}	10^{-24}	10^{-26}	10^{-28}
Rank	9,899	9,001	1,159	10	10

c	10^{-21}	10^{-22}	10^{-23}	10^{-24}	10^{-25}
Rank	9,681	9,001	6,869	1,159	10

Table 15: Robustness to gaussian additive noise.

5.4 Epsilon Effective Rank

We know that the singular values are present along the main diagonal of the matrix $\mathbf{S}$, that is obtained from the SVD of any real matrix $\mathbf{A}$. Assuming $\mathbf{S}$ to be a square matrix of dimension $n \times n$, then $s_{11}, s_{22}, .., s_{nn}$ are the singular values. The LAPACK routine that is used for the computation of SVD, conventionally, returns $\mathbf{S}$ with singular values along its diagonal sorted in decreasing order. Hence, we represent them as a sequence $(s_{11}, s_{22}, .., s_{nn})$. In the need for an alternative metric which solely depends on the singular values to calculate the rank, we define ϵ-effective rank as follows:

$$\sum_{i=1}^{k} s_{ii}^2 \geq (1-\epsilon) \sum_{i=1}^{n} s_{ii}^2 \quad \forall \epsilon \in [0,1] \tag{1}$$

The smallest value for k, to which the inequality mentioned in equation 1 holds true, is defined as the ϵ-effective rank. In words, "the smallest number of singular values whose squares sum to equal or more than $1-\epsilon$ fraction of the total sum of squares of singular values is called the ϵ-effective rank". One major advantage of this metric is that we can approximately quantify the singular value distribution (and how fast the singular values drop) when different values for ϵ are used. We repeat the same experiment for checking robustness of Press' rank to gaussian noise, but this time we also compute ϵ-effective rank for different ϵ's in addition to Press' rank to make a comparison between the two different metrics for rank calculation. The results are shown in Table 16.

c	**Press' rank**	**Effective rank** for various ϵ		
		10^{-3}	10^{-4}	10^{-5}
10^{-20}	9,899	7	10	10
10^{-21}	9,681	7	10	10
10^{-22}	9,001	7	10	10
10^{-23}	6,869	7	10	10
10^{-24}	1,159	7	10	10
10^{-25}	10	7	10	10

Table 16: Comparison between Press' rank and ϵ-effective rank.

Function	**Press' rank**	**Effective rank** for various ϵ		
		10^{-3}	10^{-4}	10^{-5}
Softmax	402	52	201	306
SS	4,979	100	297	1,038
GSS	8,989	100	559	3,456
LMS-PLIF	580	50	198	335
MoS	9,983	81	1,521	6,428

Table 17: Rank of $\mathbf{Q}_\theta$ constructed from the test set of PTB.

Function	**Press' rank**	**Effective rank** for various ϵ		
		10^{-3}	10^{-4}	10^{-5}
Softmax	402	27	141	274
SS	6,590	54	249	1,201
GSS	10,145	60	391	2,988
LMS-PLIF	513	29	150	287
MoS	15,738	49	773	5,982

Table 18: Rank of $\mathbf{Q}_\theta$ constructed from the test set of WT2.

From Table 16, we see that the use of 0.001 for ϵ seems to be on the stricter side as the actual rank of matrix under consideration is 10. Similarly, 0.0001 seems to be just right or can also be thought of being on the lenient side. Instead of using only one value for ϵ, using a couple of values starting from 0.001 and further smaller values can more or less present a clear information about the singular value distribution and the effective value of rank that is calculated from it.

We report Press' rank and ϵ-effective rank of $\mathbf{Q}_\theta$ for AWD-LSTM models with different functions trained on PTB and WT2 data sets in tables 17 and 18 respectively. The effective ranks for the ϵ values of 0.001 and 0.0001 are still low for all AWD-LSTM models with different functions, though the Press' rank is high. Therefore, we argue that Press' rank cannot be the main reason for better performance of these models on the test set.

References

Ganea, O.; Gelly, S.; Bécigneul, G.; and Severyn, A. 2019. Breaking the Softmax Bottleneck via Learnable Monotonic Pointwise Non-linearities. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97, 2073–2082. URL http://proceedings.mlr.press/v97/ganea19a.html.

Merity, S.; Keskar, N. S.; and Socher, R. 2018. Regularizing and Optimizing LSTM Language Models. *ArXiv* .

Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; and Flannery, B. P. 2007. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press.

Wang, D.; Gong, C.; and Liu, Q. 2019. Improving Neural Language Modeling via Adversarial Training. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6555–6565. Long Beach, California, USA: PMLR. URL http://proceedings.mlr.press/v97/wang19f.html.

Wang, L.; Huang, J.; Huang, K.; Hu, Z.; Wang, G.; and Gu, Q. 2020. Improving Neural Language Generation with Spectrum Control. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=ByxY8CNtvr.

Yang, Z.; Dai, Z.; Salakhutdinov, R.; and Cohen, W. W. 2017. Breaking the Softmax Bottleneck: A High-Rank RNN Language Model. *CoRR* abs/1711.03953. URL http://arxiv.org/abs/1711.03953.