
Assignment 2. Recurrent Neural Networks and Graph Neural Networks

Maurice Frank
11650656
maurice.frank@posteo.de
Code: [github](#)

1 Vanilla RNN versus LSTM

1.1 RNN derivatives

Generally we have:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{ph}} = \frac{\partial \mathcal{L}}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \mathbf{W}_{ph}} \quad (1)$$

With the two partials:

$$\frac{\partial \mathcal{L}}{\partial p_i} = - \sum_j y_j \frac{\partial \log \hat{y}_j}{\partial p_i} \quad (2)$$

$$= - \sum_j \frac{y_j}{\hat{y}_j} \frac{\partial \hat{y}_j}{\partial p_i} \quad (3)$$

$$= -y_i(1 - \hat{y}_i) - \sum_{j \neq i} \frac{y_j}{\hat{y}_j} \cdot (-\hat{y}_i \hat{y}_j) \quad (4)$$

$$= -y_i + y_i \hat{y}_i + \hat{y}_i \sum_{j \neq i} y_j \quad (5)$$

$$= \hat{y}_i \left(y_i + \sum_{j \neq i} y_j \right) - y_i \quad (6)$$

$$= \hat{y}_i - y_i \quad (7)$$

$$\iff \quad (8)$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{p}} = \mathbf{p} - \mathbf{y} \quad (9)$$

$$(10)$$

Note here it holds $\sum_i y_i = 1$ because of the one-hot encoding.

The second derivative is more direct:

$$\frac{\partial \mathbf{p}}{\partial \mathbf{W}_{ph}} = \mathbf{h}^{(T)} \quad (11)$$

$$(12)$$

leads finally to:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{ph}} = (\mathbf{p} - \mathbf{y}) \cdot \mathbf{h}^{(T)} \quad (13)$$

The derivative with respect to the hidden weight we write down in its recursive form:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{hh}} = \frac{\partial \mathcal{L}}{\partial \hat{\mathbf{y}}} \frac{\partial \hat{\mathbf{y}}}{\partial \mathbf{p}} \frac{\partial \mathbf{p}}{\partial \mathbf{h}^{(T)}} \frac{\partial \mathbf{h}^{(T)}}{\partial \mathbf{W}_{hh}} \quad (14)$$

$$\frac{\partial \mathbf{h}^{(T)}}{\partial \mathbf{W}_{hh}} = (1 - \mathbf{h}^{(T)^2}) \cdot \frac{\partial}{\partial \mathbf{W}_{hh}} \mathbf{W}_{hh} \mathbf{h}^{(T-1)} \quad (15)$$

$$= (1 - \mathbf{h}^{(T)^2}) \cdot \left[\left(\frac{\partial}{\partial \mathbf{W}_{hh}} \mathbf{W}_{hh} \right) \mathbf{h}^{(T-1)} + \mathbf{W}_{hh} \left(\frac{\partial}{\partial \mathbf{W}_{hh}} \mathbf{h}^{(T-1)} \right) \right] \quad (16)$$

$$= (1 - \mathbf{h}^{(T)^2}) \cdot \left(\mathbf{h}^{(T-1)} + \mathbf{W}_{hh} \frac{\partial \mathbf{h}^{(T-1)}}{\partial \mathbf{W}_{hh}} \right) \quad (17)$$

$$(18)$$

Because we had to write down $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{hh}}$ recursively which we do not do for $\frac{\partial \mathcal{L}}{\partial \mathbf{W}_{ph}}$ we directly see the different length of the temporal dependencies in the two computational graphs. For the hidden weights gradient we need to transverse the whole time-sequence while the other one only ever depends on the final hidden state. This gives problems in practical training probable. The partial gradients of the hidden weights might be small and they form a product. The gradient might be vanishing. In practice this might go so far to reach the numerical limits of floating point arithmetics.

1.2 Vanilla RNN code

Find the code inside `vanilla_rnn.py` and `train.py`.

1.3 Vanilla RNN experiment

See Figure 1 for a overview plot of the results and Section 1.6 for a discussion/comparison of the results.

1.4 Optimizer

SGD has problems. One of them is occurring oscillations in valleys of the loss space. SGD does not have any *memory* and thus just tries to approximate the currents face's gradient to follow down which might make the path jump around a minimum of the valley. One change to counter this problem is introducing **momentum**. Following the intuition of the physical term, the gradient with momentum gets only changed gradually not sudden in every optimizer step. This is implemented as a decaying average of gradient updates. The weights get updated as a weighted sum of the previous update and the new gradient. A second idea is to tweak the learning rate for each weight and not use a fixed η for all, yielding a **adaptive learning rate**. For those weights that change a lot (bounce around some valley) we want to reduce the update step to counteract the bouncing. This can be seen in the RMSProp [1] optimizer as described below:

$$v_t = \rho v_{t-1} + (1 - \rho) \cdot (\nabla_{\theta_t} f)^2 \quad (19)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t + \epsilon}} \cdot \nabla_{\theta_t} f \quad (20)$$

ρ defines the decaying sum. We compute the update but than divide the learning rate η for each weight by the new update. Thus oscillating weights will get a smaller update. Adam [2] optimizer works quite similar:

$$v_t = \beta_1 \cdot v_{t-1} - (1 - \beta_1) \cdot \nabla_{\theta_t} f \quad (21)$$

$$s_t = \beta_2 \cdot s_{t-1} - (1 - \beta_2) \cdot (\nabla_{\theta_t} f)^2 \quad (22)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{s_t + \epsilon}} \cdot v_t \quad (23)$$

We also adapt the learning rate per weight by dividing by the square-root of the squared gradients. But here also directly use the momentum but having the decaying sum of weight-wise gradients. β_1 and β_2 are tuneable hyperparameters.

1.5 LSTM theory

a) LSTM Gates

input modulation gate $g^{(t)}$ The input modulation gate determines candidate information from the new input (using also the old hidden state). We want our state values normalized but need also negative values (otherwise the cell values would only increase) which, as in this case, can be done with a tanh, squashing the input to $[-1, 1]$.

input gate $i^{(t)}$ The input regulates which and how much information of the input of this time step should be included in the cell and hidden state. As the input gate regulates the flow it is necessary to have its values bounded to $[0, 1]$ which can most directly achieved by squashing the values with the sigmoid.

forget gate $f^{(t)}$ The forget gate regulates which and how much information from the old cell state should be disregarded under the new information from the input (and the old hidden state). As the forget gate only changes the importance (magnitude) of the information in the cell state it should be in $[0, 1]$ which is achieved with the sigmoid.

output gate $o^{(t)}$ The output gate regulates which and how much information from the new cell state should go into the new hidden state. Again its gating the values from the other tensor which is asking for a range $[0, 1]$ achieved by the sigmoid.

b) Number of parameters

We have given $x \in \mathbb{R}^{T \times d}$ with T sequence length and d feature dimension. Further we have n hidden units. Then we have

$$4 \cdot (d \cdot n + n \cdot n + n)$$

trainable parameters for *one* LSTM cell. If we want to include the projection onto the classes c the size increases of course to:

$$4 \cdot (d \cdot n + n \cdot n + n) + n \cdot c + c$$

1.6 LSTM practice

Find the code inside `lstm.py` and `train.py`.

We train the two models (RNN and LSTM) for palindromes with sizes 5 up to 40. Both models get a hidden size of 128. As optimizer we use RMSprop with a learning rate of 0.001 until convergence of the loss. The weights for both models are initialized with He initialization [3] as it performs well compared to plain normal init.

For a overview of the results check out Figure 1. We see in general that the LSTM is able to learn the palindromes faster and for longer sequences. The RNN is only able to improve on randomness (= accuracy of 0.1) up to length 17. Further the RNN is only able to reach full accuracy for palindromes smaller than 10. The LSTM reaches perfect accuracy for all lengths but we see that for length of 23 we do not see improvement over randomness until almost 3000. If run for long enough the LSTM learns full accuracy for all tested lengths but we found that not be the interesting result here. The LSTM learning slower at higher lengths is explained in that we use the same hyperparameters for all experiments especially the learning rate. An optimization of the hyperparameters for the LSTM might speed up training for longer sequences. The experiment clearly shows that the LSTM is more capable of learning longer dependencies.

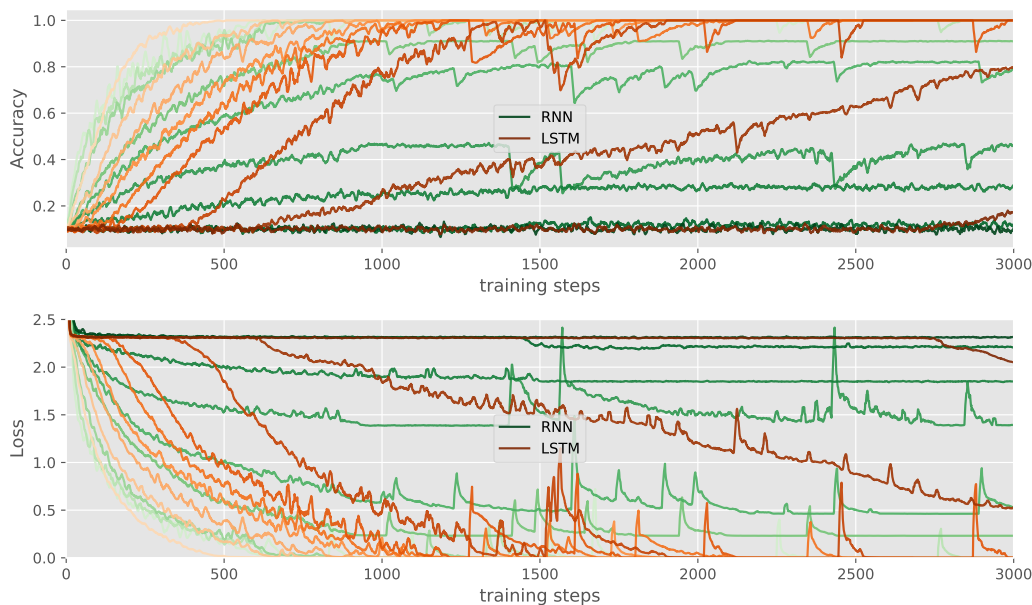


Figure 1: **Top** the accuracy and **bottom** the loss while training. Lightness of the color codes the palindrome length ranging from 5 numbers at the lightest color in steps of 2 to 23 numbers at the darkest color. (making 9 curves per model). All curves are an average of ten runs and smoothed with a box filter with width 10 for better readability. Note that two of the RNN curves never increase over 0.1 in the accuracy plot.

2 Recurrent Nets as Generative Model

2.1 Learning English from South Park and the Moses

In these experiment we train a simple character-level language model with LSTM. We train from texts from the bible and the television show South Park.

a) The Implementation

Find the code inside `model.py` and `train.py`.

We implement the basic character-level language model using two stacked LSTM cells (three for the second experiment). Both cells are followed by a Dropout unit. We use a linear projection to project the hidden state from the second LSTM down on the class predictions. As we get a sequence of characters as input we use an embedding with fixed identity weights (one-hot encoding) to process the input for the LSTM. We set the hidden size of both cells to 128. The dropout layer is set at 0.05 dropout probability.

We train the model with the Adam optimizer feeding it sequences of 30 characters from the text in mini-batches of 64. The initial learning rate is set at $2e-3$. Further we decay the global learning rate every 5000 steps by 0.04.

We preprocess the text files by reducing the set of characters. All text is made lowercase and line breaks are removed. We remove special, uncommon characters (e.g. acutes from German and Spanish) replacing them: e.g. $\ddot{o} \rightarrow o$ and $;\rightarrow ,$.

b) The results

For training we feed the model with all transcripts of the television series **South Park**. In total the transcript have 813k words. See Figure 2 for the training progress in the first $1e5$ steps.

To understand the generative power of our model qualitatively we sample from the language model at different steps during training. Sampling is done by picking a first character uniformly by random. Next we iteratively feed the character into the model generate new hidden and cell states. From the hidden state we can sample a new character under different temperatures. The new character is saved in the string and fed back to the model with the new hidden and cell state. We do this until we reached the desired output length.

Following are text samples of length 20, 30, 50 and 120 at step 0, 1e4, 2e4, 3e4, 4e4, 5e4, 1e6, 2e6 and 5e6. Each step we sample at temperatures $T \in 0, 0.5, 1$ and 2.0 . Note that temperatures 0 corresponds to greedy sampling, meaning we pick the character with the highest probability.

0			10000			20000		
Temp	l	Samples	Temp	l	Samples	Temp	l	Samples
0.0	20	riiiiiiiiiiiiiiiii	0.0	20	me to the stupid stu	0.0	20	9 the stupid people
	30	5iiiiiiiiiiiiiiiiiiii		30	re to the strange of the stupi		30	and the stupid asshole! wh
	50	äiiiiiiiiiiiiiiiiiiii		50	that's not the power to the boys are go		50	g and the principal problem with the stu
0.5	120	iiiiiiiiiiiiiiiiiiii	0.5	120	nna be a	0.5	120	dents and
		oiiiiiiiiiiiiiiiiiiii			ze is the problem than the problem to th			can that the students and the students a
		iiiiiiiiiiiiiiiiiiii			e popped to the balls are gonna be so th			nd the police and the police and started
1.0	120	iiiiiiiiiiiiiiiiiiii	1.0	120	e problem that the problem of the part o	1.0	120	to the police and state the part of the
2.0	20	yjw2b3b6 nd. w,m54tw	2.0	20	! i know i thought i	2.0	20	-hold on. god damni
	30	7-edbx2)dg4n7syam enlahvã sx-		30	11 and sooo and then we can be		30) that's what i don't have a
	50	r8'/du1c6qzc1si7.lklsmwoui"xznã68-a8gg z		50	got a spilitory. the hell are you beauti		50	ou want to show the poor more than your
1.0	120	w'e)!/pb6	1.0	120	ful his mo	1.0	120	problem st
		kee,e)j8noza,a/6enra""h'f'a79d'0gãu'ya			/fuck your son in your new problem and i			5 from here for here! we can do it for
		!nr-(opt,s(t ode)!,hz3z9(sc'jq/b zpr5ãc			'm sure your looks like the lately his n			a boy is for a bad of the planet. what a
2.0	120	txlm9uid(bnd9zr6/p ãlnpegim nzxh ?7nskc	2.0	120	ight to the bame! and we have to play wi	2.0	120	re we gonna do it out! what are we gonna
0.0	20	el28jn25(1815yzwthr3	0.0	20	nrish are dellagin's	0.0	20	tle son is so any ne
	30	0..fg'(iaa 9'ãxã)d6a 8li7oabaf		30	hiltor. alto si-.ããdisolant s		30	quut up and shows! we want it?
	50	ptnctn,n,h k.-6ybi1b1/,mdf5a(8xenxmmu7bsx		50	quacking are on mandayout kike the burg.		50	fell is whatever if here in the stuff th
0.5	120	du3lvt95b9	0.5	120	that's? k	0.5	120	at we coul
		rx54 ?5taif'cs1c)la36wsy0'2eodj6/?y.5yq?			lor place! all what we've reading much-b			! grabba basky with a news. no, kyle. ma
		q/2 ocy5,(ctzq'1n?sojh38ahgp() 9m3.p9lhs			leh'is is brolly hangh! keep you! i'm ju			ybe you talk to you that? i don't know y
1.0	120	7918h3lã1ãf1"0.ñ3s13ve zk'ã,-u85(tt dp	1.0	120	s gonna kill by a hot. .r. tom. yeah! h	1.0	120	our veal. my pockets, zammate your legra
2.0	20	ar8ld7gb7nls5pzxj 5f	2.0	20	yed. yoh,! whaa, wha	2.0	20	8 totarnap quiete ch
	30	ufnjur,fnc2o)h9hy?r9.fkpn8)8wv		30	, huh8mu, nsy! oh what,!, 'yeaw		30	(clein! lond.,ncomaa-lopeous c
	50	jkw(-unz1vm'n3'ã/?r6rz(le0ã c8q k0(pnff		50	yetriuze be bocy toks. ugs! 4" - you ju		50	p, "" nice, nobody! g'ee!! yemph.) o-how
1.0	120	4.oy7)san3	1.0	120	st lose vi	1.0	120	seefo elea
		?-pi!'8l6)n-qj/50.j.naq7lebvnmlã" c02mj			(122.ã peofjies, bervltsalyfatquemasmine,			2l...lu-yuwwakkih? caught. ""n'tcufort, d
		0t9 vekhj-7zbw-8f6f-dzq?-m.(q7z'73s)7-u			you'do've fwire 69fven cull,! bye.!!).			nivo weggghy newra! geraldon(ensides you
2.0	120	rn eqtã"nd2omjji8 (7j30xylc9ore.ubon762i	2.0	120	me! it's ville, ""days, sv.!l poft one	2.0	120	know! in-carv twa! (dudd-k egerpizpsnc.
30000			40000			50000		
Temp	l	Samples	Temp	l	Samples	Temp	l	Samples
0.0	20	per the problems and	0.0	20	the stupid ass of t	0.0	20	- and then i can se
	30	e the problem is a little boys		30	the stuff in the students and		30	f the students are so that the
	50	n the protect the probably in the contro		50	you guys are gonna be a children and sta		50	ing to see you all the start on the part
0.5	120	l of the p	0.5	120	y out of h	0.5	120	y of the p
		- and i don't know what i was the paren			s and the problem that the state of the			re a little boy in the principal stupid
		ts are the stuff and i was the problem t			students and the problem that the studen			asshole! what the hell are you doing? i
1.0	120	han the park country in the state to the	1.0	120	ts and start a second of the stupid ass	1.0	120	don't know what i can do it all the stu
2.0	20	? oh my god! well,	2.0	20	ll you have a drink	2.0	20	good, the only perso
	30	g all the kids hate your vampi		30	t that that we got reality who		30	10 people who present. you don
	50	/friends to put the bus. well, sure is a		50	t his son's friends and a second. you gu		50	2 mm. cartman, you're not going to make
1.0	120	little fu	1.0	120	ys take a	1.0	120	them bette
		day cartman! it sure is true. he's kille			call my house. that thing is stoppin' he			s on the people be a little boy the stat
		d, and i want to do is what you understa			r canadian for a time! ...and i know ho			us of her there. tell you in here. i can
2.0	120	nd that there was my friends and some ho	2.0	120	w you wanna do this and well then i have	2.0	120	't take it anymore was some construction
0.0	20	me act now. that's w	0.0	20	ãdreambit careful, i	0.0	20	okay. 'cause you com
	30	give me, even going up with my		30	xice how me a mall eating. tha		30	ll i've... good way, and sweet
	50	2s? i don't know a people have a lot of		50	ball maybe insured on our odmedent. it'		50	7 friends off, too. my legs of force her
0.5	120	words agai	0.5	120	s not on t	0.5	120	e he give
		5 down't you guys are not all the show t			aking a bunch about you! alright. oh, c			pash! you decighcion the poxauld first
		ogety goddamned doctor is a crow, butter			ows! that's nice back that does. and uh			comes in trualing that saying this time.
1.0	120	s! that does with ana good oraig, where'	1.0	120	yeah, it'll get the titzolisys. six stor	1.0	120	oh yeah. okay kids, donleya! ned genefi
2.0	20	n aypofa? yeah!!) u	2.0	20	-e! your habsy. eric	2.0	20	u don't know. ãriva
	30	quedsplove., not theared lovin		30	7395.""- stapa.nlona ded, you		30	julth not an/mh??pkmmn! holy,
	50	yw. xlezs, v mrtol! govertuh you ya! um r		50	iaand quest monmils, eric. moutupt a!!! t		50	tyal! we!sex, subbeld's bad alright-,i
1.0	120	test. hyeh	1.0	120	hanks, cla	1.0	120	mul! bemp
		6e-c6? ma'poghutlike,"waitbmecy. a adulc			ah'-'! ...ohrs, who duid "c?ãcalmndy!) ta			?, -ããuh i camebokalo's?!iravy?! oh, s
		cuma" zey ninety, 3.6 wql-fs' woman?!ãk			?! uh tox into ys-ave. shho! o-kay! you,			top!-uw, all, uuuf-if there suddes8 her?
2.0	120	yre? thh). wamnasninqn 4) i-cake it gly	2.0	120	9, quickly. (kittle koek!!r me cy were	2.0	120	unfniyor mado. start! afriewt.! your l

100000			200000			500000		
Temp	l	Samples	Temp	l	Samples	Temp	l	Samples
0.0	20	f the stupid asshole	0.0	20	xt to the bathroom a	0.0	20	x and then i can see
	30	we have to get the star really		30	40000 per cent of the state of		30	, i want to be a company that
	50	u guys are the only thing to do it there and i don		50	(a friend is a little bit of a bitch! w hat the he		50	quite a lot of money to see the start a lot of mor
0.5	120	80 per cent to the soul of the state and so the school stay away from the countr y is a big deal with the start a second	0.5	120	n the stupid party of the store for the second that the world is a little boy wh o was the problem. what are you doing?	0.5	120	5 month. what are you doing? we have to do it the students are the one that the world is a little bit of the problem.
	20	ke the pain! we want		20	85 dollars signing t		20	n i don't feel so m
	30	2000, i'm a big and i can't do		30	ing the town of a super be sat		30	3 dollars. i love you guys, i
1.0	50	ou guys! yeah, we have to work at my bod y. i don't	1.0	50	00 per cent of them. hello eric, that's it there.	1.0	50	/11 has pretty good to go one different boys and f
	120	we have to kill you six hours. they did it! the things to get a really seeing t he one things is. we are the brains and		120	ke in a son is about to be a big deal to god is having a new town of all the red life in sandwich thing? so for you to s		120	2002! he was the united states with the armas place in the fish and that's okay. all right, we are so funny! we can do i
	20	ht at us saw... all		20	for that idea. are	1.0	20	w, that noticed you!
2.0	30	? don't they come a meters! wh	2.0	30) doh, if a scott. i know we'		30	- she's too... ruuuurribuy! t
	50	, you know what they aren't any of that guy travel		50) your sriss of mouths of charge of you de by him		50	10,000 one to this out of hero...you'll meet a daw
	120	? really? you have to be costaiint. so, nuh dougless is the confederated chinpo kes you dope? could you land it... sandw		120	r you now bitching my part of my towelie !! i know it here? look, we will cut off my balls, red measure?! get your grandm		120	gonna build your flute boy in my olved q uit. go sure. it's serious! every piss, huh?! oh, hey, changedage the stupid au
2.0	20	, ver disk thos girl	2.0	20	goaj. (go we'll bo,	2.0	20).qh avoid theng "h
	30	(k. graswmak!! hs, ly- wen'- n		30	evweac eated itge-joren l.? i		30	- ad zumir's reevc dlamlant! l
	50	4 ! eric cat, kyloh, -mosely!/-sssuper ing,s waal		50	've gravi my vacco., "s?? kewnu , ms. ng -ctdai.,mi		50	ub! yeephclexto, cartman, the run. o,...) boy. ste
2.0	120	? bopha nicpdorations, a chinaple th you r-rect kickin'cybun tripi and 'cd'of-)-whh ibz. yem.y?) oh... with?. (hm.!) if	2.0	120	872!0 yuk. ?que? whe- hey mirgith rouetn put yoba goddami-engal"" go back with 3/ 5. obceaoab, ike, 'f's lots autlakffful i	2.0	120	- get mm."" nribly dude! p-juros.!!fiz e i bb gallawes—c-curity!"" caren sogar th! right ninleys, garyeow? i'm i?" ...

We see some general trends. At temperature of 2.0 the model never able to form words which suggests that this is too high. Most interesting we see the model being able to generate almost coherent sentence early on in training as e.g. at step 1000 with $T = 0.5$.

In a second experiment we feed the network the **five books of Moses** from the old testament in the version of King James. Compared to the previous transcript this text only has 156k words. We extend the model architecture to check if that increases the generative power. For that we us *three* stacked LSTM cells. Further we feed sequences of 40 characters and decay the global learning rate only all 20k steps. All other parameters are set as before.

Trainings progress was similar to the first experiment so we just give exemplary qualitative results below. All examples are at temperature $T = 0.5$:

Step	Sample sentence (first letter random) $T = 0.5$
1000	d the proord of the make of the land the panctoure at the lord, and his hase unth alt the come of the thald the lord the
5000	pearn, and the place of the sight of the children of israel, and put the place of the children of israel, and the place
7000	and say unto moses, and keep the tabernacle, and that the lord thy god in the mighty will the lord thy god
50000	had sinners. and the lord spake unto moses, saying, speak unto the children of israel in my hand, and thy sons in the
150000	me a graven image, or a meat offering: one golden spoon of ten shekels, full of incense, and the hittites

Lastly we can use this language model to complete sentences. For that we take the cleaned input and also feed it to the model character by character. After going through the part sentence we can as before continue sampling new characters:

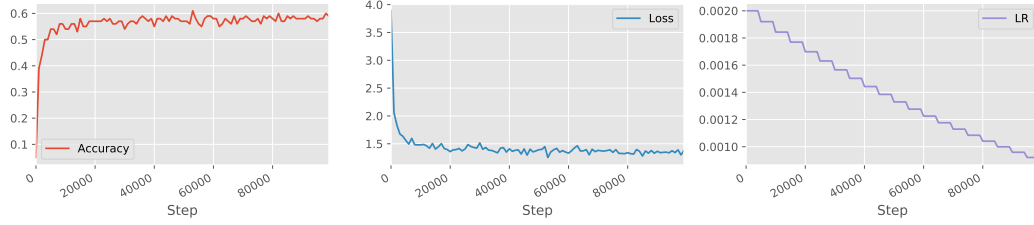


Figure 2: Accuracy, Loss and the decaying Learning Rate for the LSTM language model during training using the South Path transcripts. Note that accuracy and loss are a bad surrogate for our task as they describe the capability of the model to predict the next character given a string while our actual goal is to get a generative model.

Beginning	Hallucinated completion
you shall. be burning, and say unto them, when ye come unto the lord for the presence of the people from the morning, and the priest shall burn the thing that good land.
god will. proclaim the land which the lord thy god shall choose to place his name there. and israel dwelt in the seed of the land, and the lord hath destroyed them.
in the beginning god. created his sight, and the land which the lord hath bound her son. and the lord said unto moyses, the lord had said unto them, they shall not be destroyed in the field
yoshua bengio is. god, and the perizzites, and the two tables of the covenant with them. and the lord said unto moyses, i will put them on the earth.
schmidhuber is. a meat offering, and the other side of the tabernacle of the congregation, and when it is unclean unto you.

3 Graph Neural Networks

3.1 Forward Layer

a) Message passing in a GNN

The \hat{A} matrix contains the edge information between the nodes in the graph. This includes the self-connections of all nodes from the identity $\mathbb{1}_N$. We update the activations with $H^{l+1} = \sigma(\hat{A}H^{(l)}W^{(l)})$. So we update the activation associated with one node only for the activations of edges that have non-zero edge weights in \hat{A} . A node is changed by the nodes its connected to. In reverse we can see that information in one node can propagate to all connected adjacent nodes in one time step which can be visualized as the message passing over the graph.

b) How many layers to propagate a message along three hops?

In every layer of the GCN you can propagate information along one edge so to let information reach a node three hops away we would need three layers.

3.2 Applications of GNN

In *Multi-Granularity Reasoning for Social Relation Recognition from Images* [4] the authors build graphs on images of humans. The graphs describe first the relationship of a person in the image with its surrounding objects amongst other things other persons in the image. A second graph represents the pose of each person in the image. On these two types of graphs they use Graph Convolutional Networks (GCN) to predict the social relationship of these persons. For example in an image of parent and child the bending pose of the parent and

the connection of the two persons in the Person-Object graph lets the method infer their relationship.

Next in *Disease Prediction using Graph Convolutional Networks: Application to Autism Spectrum Disorder and Alzheimer's Disease* [5] propose to use GCN for medical image processing. The nodes of the graph in this setting are features of medical image acquisitions which in their experiments are gathered from structural and functional MRI. As they predict the health state of multiple individuals at once the graph consists of many image features of multiple persons. The edges here describe the phenotypical similarity between two individuals which are described by categorical medical data (e.g. sex). The GCN uses this graph to predict the health state of the population.

Lastly in *Temporal Relational Ranking for Stock Prediction* [6] we seen an application of GCN in stock prediction. Here the nodes are features capturing the historical information of one stock and the edges capture the relations between two companies stock. Both historical intra and inter stock features are generated by LSTM from historical stock data. Again we build a graph from the edges and nodes and use GCN to predict in this case the stocks next day performance at the market.

3.3 Comparison and Combination of GNN and RNN

a) Comparing GNN and RNN

Using a RNN assumes our data is orderable along one axis, meaning every node of information is at most connected to one *previous* and one *next* node of information and that no two nodes have the same previous or next node. For the GNN we assume data whose node of information is related to an irregular number of other nodes. One major difference for the two is that the GNN can give edges between, relationship between two states, a weight or a feature itself. More specific for the RNN we assume a equidistant (temporal etc.) edge between each node while for the graph we can adjust this.

RNN is more suitable and will outperform GNN if this rigid structure is fulfilled. For example learning word embeddings (words are sequences of letters with no given edge differences) are still a successful application of RNN.

One can assume GNN outperforming an RNN model if working on structured data that warrants graph representation. Further many highly regular structured datasets can be preprocessed to a form in which a GNN can be fully utilized. For example texts are themselves not directly a graph representation but if we preprocess them to get find syntactic structure and maybe dependency graphs the application of GNN becomes more obvious.

One topic were GNN are the useable and RNN models miss the complexity to capture the data's complexity is in Physics and Chemistry [7]. One example here is molecular fingerprinting. The highly complex graph structure of molecules makes it difficult to embed this structure through a RNN model but there are multiple works on using GNN or more specific GCN, for working with molecule data.

b) Combining GNN and RNN

We saw such a combination of RNN and GNN models in Section 3.2 already. The work on prediction of stock markets used LSTM and GCN in conjunction. In there each node of graph related to a linear temporal structure, the historic sock data. As those linear temporal graphs where themselves related to each other one can use a graph model to analyze.

The reverse idea also seems plausible by using a RNN to go over nodes of graphs. Imaginable might here be an application where the data is organized in a graph, let's say a social graph, but is subject to temporal changes. A GNN could be used to analyze the state of the graph at one given point while the RNN uses these temporal states and can for example predict future states from that.

References

- [1] Geoffrey Hinton. Lecture 6 Overview of mini-batch gradient descent. URL http://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.
- [2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization.
- [3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. URL <http://arxiv.org/abs/1502.01852>.
- [4] Meng Zhang, Xincheng Liu, Wu Liu, Anfu Zhou, Huadong Ma, and Tao Mei. Multi-Granularity Reasoning for Social Relation Recognition from Images. URL <http://arxiv.org/abs/1901.03067>.
- [5] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease Prediction using Graph Convolutional Networks: Application to Autism Spectrum Disorder and Alzheimer's Disease. 48:117–130. ISSN 13618415. doi: 10.1016/j.media.2018.06.001. URL <http://arxiv.org/abs/1806.01738>.
- [6] Fuli Feng, Xiangnan He, Xiang Wang, Cheng Luo, Yiqun Liu, and Tat-Seng Chua. Temporal Relational Ranking for Stock Prediction. 37(2):27:1–27:30. ISSN 1046-8188. doi: 10.1145/3309547. URL <http://doi.acm.org/10.1145/3309547>.
- [7] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph Neural Networks: A Review of Methods and Applications. URL <http://arxiv.org/abs/1812.08434>.