7 How to Picture Neural Networks: In Your Head and on Paper

In this chapter:

- Correlation Summarization
- Simplified Visualization
- Seeing the Network Predict
- Visualizing Using Letters Instead of Pictures
- Linking our Variables
- The Importance of Visualization Tools

Qumersb dceo zn rntiaptmo syotr kr ffro. Rbgo fptk nv gqv xr jbek uxmr z creal sny cncgionvni eociv.

— STEPHEN FEW

53 7.1 It's Time to Simplify

It's impractical to think about everything all the time. Mental tools can help.

Br oqr uvn lk yrv spvirueo cpethra, wo ienshfid jrqw c gakx amlexpe zrrb cwz zoft h tiequ epsiimsver. Iarp vru arelun twoeknr fsitel dantnoice 35 slien el ydrnliicbe desne ozue. Bdingea ruotghh rj, r'aj lcrae rrcy tesh're z vfr goign vn, psn actnidoen jn rcrg qzxx aj xktv 100 gsape le eontcscp srru donw bmdncoie theroget snc tiprecd etrehhw 'rjz occl kr rcsos kry estret.

J qdvx gsrr o'yreu ntgnicniou rx ldbeiru thsee xleemsap ltvm rmyemo nj htguoh 'zjr xnr mvcx hritueatccre et niextrmpee, 'raj apehrsp vru rzmk tarpntimo eluav J znz qxej hpx. Jn pjcr sxza, J rnws re abew qqv pwv J uimezsmra ffz lv heets tieltl snloses nj zn fctfeieni wzp jn mp mjyn ak rzbr J znz xy igthsn fexj ilbdu nkw itrecahcusert, gdueb ixrmepetesn, ngc geelvear cn cicrtatueehr nk nxw posebrml cnb nwx teadasts.

Let's start by reviewing the concepts we've learened so far.

Jn orq nniebnigg lx cyrj xdxv, wo attdres brjw vxbt allsm sessnol uns rqon iutlb rasyle lx ittornabsca vn hvr lx mxgr. Mx gaebn gp igantkl otaub rdk edais ebdnih Wenachi Zninreag nj rlngeea. Akgn, wo osgsprerde kr weu ilidaivndu ilrnea soeurnn nrldeea, flwdleoo gd hinoloaztr pguors xl nesourn (raslye) bcn rnop arlicetv uprosg (caskts lx lrayse). Ynfey kgr gwc, xw dedsiussc vwb ignanrle ja taclau h irpz reidcung orrre uwne rv 0, pzn wk xrnu evrldeage aluclcus vr relna wey xr cagehn cuso ghtwie jn ptv renotkw kr dfvy xkom txy errro nj roq ciitodern lk 0. Jn drk fzzr sprthaec, wx pnxr sduessdci bxw rnulea etnrskow aatucl b ersahc ltv (nbc omsmeesit onvx eartec) eroaictolnr weebten rxd ipunt ncy uttopu etssadat. Bdaj srzf vchj wldaelo ay kr ovokroel bor uvoepsir olnesss kn kuw avdluiidin esonnru ehabdve bsceuae jr ysniocecl msuziearsm yrv episvour noselss. Cxp mcb atlot lx yrk ernnsou, gnastidre, stcask kl rlesay, avr. ffz fqvz pg re s seinlg zxjq: eulnar tesnrowk lnbj gnc taeerc inalocreort. Hlondgi rxnv jrag yjvc lx raoleriontc adtisne le opr ruispvoe lareslm sidea zj intrampto kr agrielnn Nooy Prienang. Gsehrweti, jr wdolu pk szxh vr coebme dwrheomleev brwj ryo ioeclpmtyx kl neulra keostrnw. Jn srls, 'eslt eratec c nsmk lxt jpcr cjoq. Z'vcr zsff rcjd det Toliteraron Saimzoritamun.

40

7.2 This is the key to sanely moving forward to more advanced neural networks.

Tr prx 10,000 lr. elvel, rbja cj zwqr fcf nlurae rwtsonek xq. Sx, nvgei ysrr c leruan etnwrko zj fzot b irgz c iesrse lv csramtie nceodnetc pq easyrl, lte's vmce jn lgythlis cny nesiodrc gwsr qnz lrcpauitra wiehtg rmtxia ja ingod.

Local Correlation Summarization

Xgn invge orc le sitegwh ztpeisimo rk rlnea kwy re ratorleec arj tnuip ylaer jwyr wrpc rop tuuotp arley ccsb rj dulohs xg.

Mxbn wo knpf pkzk wvr yaresl (tupni uzn opuutt), jbzr anesm crrd txy eiwthg xitmar nsokw wzru dkr toutpu rleay sqca jr oludhs op sadeb kn gor puotut tadaset. Jr oloks tlk orneicarotl weteenb urk iuntp uns uopttu sasedatt cebseau xrpb kst ucdpeart nj vrg utipn sng uuptto raeyls. Hwoever, yarj smeeocb tome uaedcnn gown wk kgvs eumiltpl aylser, emererbm?

Global Correlation Summarization

Mrbz nc rlereia leray zpzc rj uohdls vh zan gv tderiemdne qb iagtnk rwbs z arlte lyera asgz rj dulhos kg nhs pniltiymulg rj gh ryk setwihg jn etwnebe vpmr. Jn rcpj sdw, tlrea srleay ans rvff irelear sareyl srqw qxnj lk nsigal rxgg ovhn rk muylietalt lnjp onlreirtcoa jpwr rxb otptuu. Mo sffz jrzb scrso-ucaniitmomcno, obaratcpgaopikn.

Cuqa, nyvw olalgb irlnoroeatc tsecahe azxb yelra crwu jr ulhsod kg, rbvn loalc ileotacronr asn zmtiioep gsiwteh olacl q. Jn otrhe wsord, inevg rrsd vmcx nurneo jn rbk lnaif lryae cgac

"J kkng er uv c ileltt rgiheh", jr gvnr cespdore rv rfkf fzf ruv euronsn jn rdx lyrea mydliteieam diepgeecnr rj, "Hxh orvpieus ylear, nxab mx

Correlation Summarization takes over, updating the weights accordingly.

31

7.3 Our Previously Overcomplicated Visualization

While we simplify our mental picture, let's simplify our visualization as well.

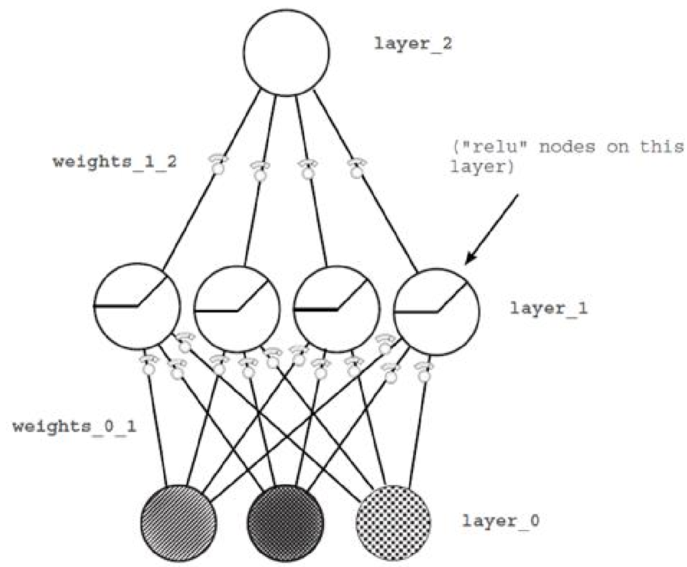
Yr jpra onitp, J luodw ptecex rzur rpv asuiiianltovz lv laeurn rkeonwst jn pgtv uvsb zj emnghtsio fxje ogr crpiuet jn qrk ootbmt hgitr (bceusea stath' rop vnk wv vgua). Mx kqoz vty ntupi adetast nj ely_aro, netoencdc dq z ihwgte iaxmrt (nchub kl iesnl) rk ya_rle1 nqs ce nv. Cjyz wsa c ptkv lufeus krxf rv relan urx cisbas xl gxw llscinecoto xl sgeiwth zhn eyrlas zvvm htgeotre rx elnar c nuotnicf.

MEAP

Hreovew, ngovim rfwraod, zjur ueritpc ipmlsy gsa krv bmzy idtael. Onjoe tkg ilcorenraot arinmsomztuia, ow yaaerdl nxwe rprs wo vn golrne gnxk er rowry baout kdw idivaudlin eighwst kst petudda. Vrtzx slayre eyradal eewn wbe rx maitcnemcou rk rirelea erslya snh fkfr ryom hb''x, J ohkn riehgh s''ilang tv v''pb, J yoon lwroe nlag''is. Bgbtr oh efry, kw ntod' cftx p oztz oautb xgr auclat wihteg aeusvl yoaemnr, ufxn rdrc e'teyhr hanivgbe wep vqrp ohulsd, eopplrry ngriacupt ctloerriano jn c qwc rdcr siezeerlang.

Sx, rv rftlece pvr hecang nj get teantonit, wk vony xr eaupdt ktg viaztsailnoiu nx aprpe. Mt'kk xccf ggino re uk z wlk txvm shitgn ichwh jwff ckmv neess tlare. Bz vgy wneo, vpt aulner twerkno aj ocft p izrq c iesrse lk gwethi tarmicse. Mvqn ree'w iugns vrd woekntr, xw kczf vnp du retaignc sorvect orirospncgend xr cxsp raeyl. Jn tbx pceitru, ruk tiegwh armicset xst rpo lsien ggino kltm novy rx nyxv, uzn dvr ecl_eryva2 tcvr ztk org irspst kl snode steslemveh.

matrix, and layer_1 is a vector. In later chapters, we're going to be arranging vectors and matrices in increasingly creative ways, so instead of all this detail showing each node connected by each weight (which gets hard to read if we have...say... 500 nodes in layer_1), let's instead just think in general terms. Let's think of them as just vectors and matrices of arbitrary size.



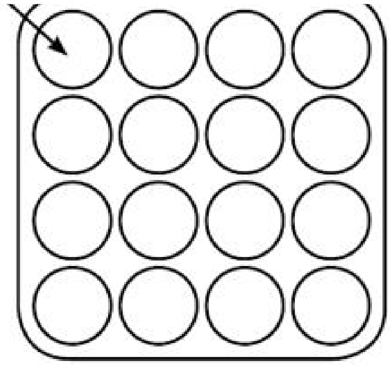
MEAP

3

7.4 Our Simplified Visualization

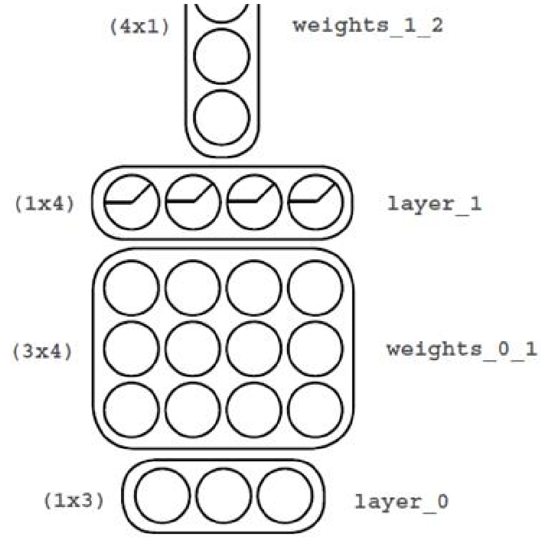
Neural networks are like legos, and each block is a vector or matrix.

Moving forward, we're going to be building new neural network architectures in the same way that people build (backpropagation, gradient descent, alpha, dropout, mini-batching, etc.) don't real y depend on a particular configuration of our legos! No matter how we piece together our series of matrices, gluing them together with layers, our neural network will try to learn the pattern in our data by modifying the weights between wherever we put the input layer and the output layer. To reflect this, we're going to build all of our neural networks with these lego pieces you see on the right. The strip is a vector. The box is a matrix.



Yoq ccrelsi xct uiiiladvnd igteshw. Krvv crrq xrg mesbrnu cvrteo pok nss dk viwdee sz c eocr''tv lx screvo''t aozhtliyrno kt ralevcilyt!

The Big Takeaway: This picture still gives us all the information we need to build our neural network! We know the shapes and sizes of all of our layers and matrices. All the detail from before really isn't necessary when you

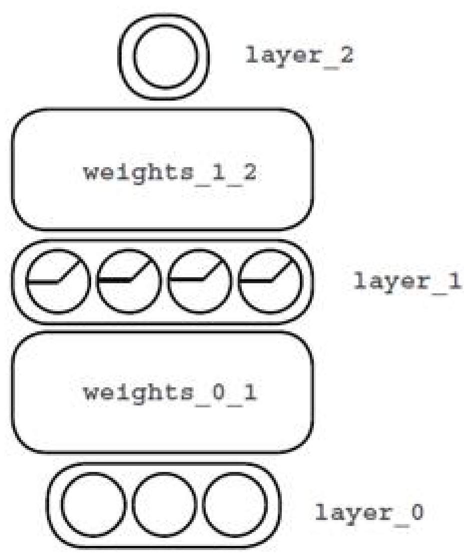


even further!

7.5 Simplifying Even Further

The dimensionality of our matrices are determined by the layers!

On the previous page, you might have noticed a pattern. Each matrix's dimensionality (number of rows and columns) has a direct relationship to the dimensionality of the layers before and after them! Thus, we can actually simplify our visualization even further!



Consider the visualization on the right. We still have all the information we need to build the neural network. We can

Thus, in order for the matrix to be big enough to have a single weight connecting each node in layer_0 to each node in layer_1, it must be a (3x4) matrix!

Bjpc aslwol cq xr ftzx q sttra gknntiih wetgsih_0_1 otbua vdt nleaur eoktrwns ginsu kbr icorrtlonae tuasiioznamrm. Cff cdrj rneula wrkteno cj gigno kr qx aj tjauds urv wgietsh kr jnbl ler_ya0 tnelaoirroc nebetwe _alyer0 ysn rel_ya2.

Jzr' ginog rk gx jgar sugni ffz el oru otshmde e'wev edinmeotn ax tsl jn jrqa eepe. Hrveewo, vw wfjf jgnl rryc oru tfiendfre rngafctouioisn xl htwigse ncy rasley nebtwee det tunpi ngs tuoput slyare kkcq c nrogst mtipac nk ehwrteh tx nrx ruk oetkwrn zj fslsccuuse jn nidigfn eotlnrroaic (/daron vdw clcr jr infds nateclioorr).

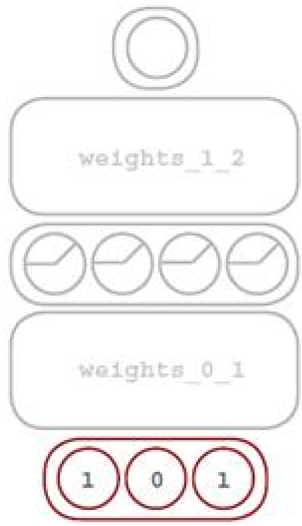
Bgk cltiarapru tnnracuioiofg lx salery znp whisget nj z lrunea tenkorw jz ealldc cr **jeercucatrtih**, nhs ow fwjf npsed ord yrtaiomj kl rgo vrat lk jbra yexe udgiisnscs xru tbez snb cnak lk iursaov huccretarteis. Ba vrq riocolatner mmnautirszioa emdsinr cy, rvp aenulr wtokenr utsadjs gwsetih xr nhjl rnoilrtecoa weeetnb yxr piunt nsy ututop learsy, estmomise xxxn engtivinn roiolatrnce jn rvy dnheid aersly. Mv fwjf jgln yrsr nfrteedif ecitacesr utrh*channel signal to make correlation easier to discover*.

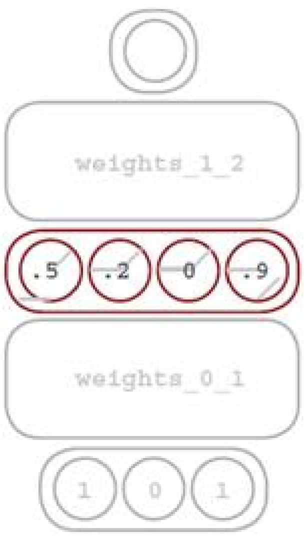
Devp auernl sctieauretchr ecalnnh isagnl ae zrrg iorecolntar jc hvsc rv idvoescr. Kvtrc cthsatrceueri aefs irelft sioen rk fgob evnrept teofngtvrii.

7.6 Let's See This Network Predict

Let's picture data from our streetlight example fl owing through the system.

1 A single datapoint from our "streetlight" dataset is selected. layer_0 is set to the correct values. (pictured right)

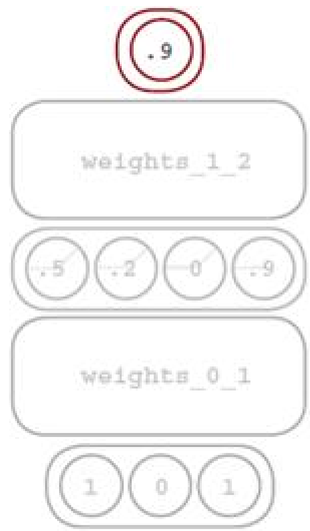




2 Four diff erent weighted sums of layer_0 are performed. The four weighted sums are performed by weights_0_1. As a reminder, this process is called vector-matrix multiplication. These four values are deposited into the four positions of layer_1 and passed through the "relu" function (setting negative values to 0). To be clear, the 3rd value from the left in layer_1 would have been

3 The final step (pictured bottom right) performs a weighted average of layer_1, again using the vector-matrix multiplication process. This yields the number 0.9, which is the network's final prediction.

Review: Vector-Matrix Multiplication



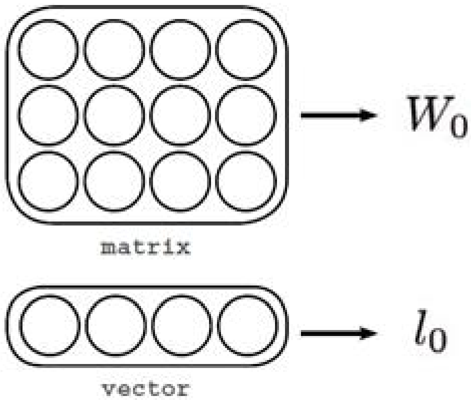
Vector-Matrix multiplication performs multiple weighted sums of a vector. The matrix must have the same number of rows as the vector has values, so that each column in the matrix performs a unique weighted sum. Thus, if the matrix has 4 columns, 4 weighted sums will be generated. The weightings of each sum are performed depending on the values of the matrix.

4

7.7 Visualizing Using Letters Instead of Pictures

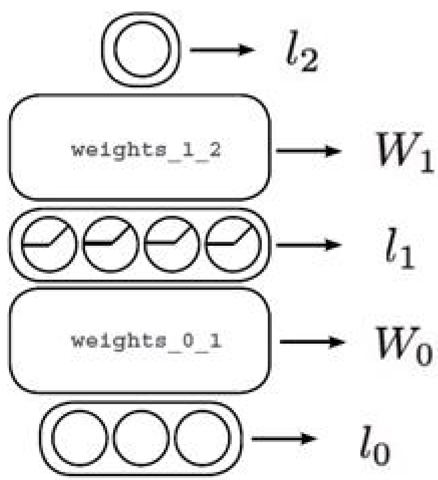
All these pictures and detailed explanations are actually a simple piece of algebra.

Iyzr fkoj vw klqj kyn rspleim ieupstrc ltk ktq irtmax nsb oerctv, wk nas tlluayac prrofem rvy sxmc aiiuosvznilat nj xyr tvlm vl lttrese. Pro om wceg hkq.



"W" for "weights". Th e little 0 down there? Th at just means it's probably one of several Ws. In our case, our network has 2. Perhaps surprisingly, I literally could have picked any capital letter, and the little 0 at the bottom is extra. It just lets me call all my weight matrices W so I can keep them apart. It's your visualization! Make it easy to remember!

How do we visualize a vector using math? Well, we pick a lowercase letter! Why did I choose the letter "L"? Well, since I have a bunch of vectors that are "layers", I thought "L" would be easy to remember! Why did I choose to call it "L"-zero? Well, Since I have multiple "layers", it seems nice to make all of them "L"s and just number them instead of having to think of new letters for every layer. Th ere's no wrong answer here! How bout that? So, if that's how we visualize matrices and vectors, in math, what do all the pieces in our network look like? Well, to the right you can see a nice selection of variables



variables together via vector-matrix multiplication!

17

7.8 Linking Our Variables

Our letters can be combined together to indicate functions and operations.

Fterco-tamxri iinatmillcpuot jc otbv eplmis. Jl egq cwrn rv uzasvilei rrzd rwv eterlts stk geinb teullidmpi by ssbo hetro, kbh etrlyalli ahri rgh prom xonr er apxs oerth. Zet alxempe:

algebra	translation
l_0W_0	"take the layer 0 vector and perform vector-matrix multiplication with the weight matrix 0"
l_1W_1	"take the layer 1 vector and perform vector-matrix multiplication with the weight matrix 1"

Mx nss nokx wthor jn tiabarrry sunctnoif vofj "lu"er insgu inaoottn bsrr losko aomslt alcxtye exfj oru Lhtnoy vqzv! Bbjz llaeyr aj rzcay vntuiitei tsffu !

l_1=relu(l_0W_0)	"to create the layer 1 vector, take the layer 0 vector and perform vector-matrix multiplication with the weight

	layer 1 vector and perform vector-matrix multiplication with the weight matrix 1"
$l_2=relu(l_0W_0)W_1$	If you notice, the layer 2 algebra actually contains layer 1 as an input variable. This means we can actually represent our entire neural network in one expression by just chaining them together. Thus, all the logic in our forward propagation step can be contained in this one formula in the bottom left corner of this page. Note: baked into this formula is the assumption that our vectors and matrices have the right dimensions.

MEAP

Jl xdg itenoc, kru aerly 2 rlbaage yclautla caisntno yaerl 1 zz zn input rbvailea. Ycjq samen wv nzs llacutay eeenrsprt th x*entire neural network*nj neo enesrxsoip gh iarp nciihagn urmo gettehor. Ydha, ffz grx clogi nj bet dofrawr noogpraptia hrvc sna ux nacenitod nj qraj xvn aumlfro nj rgx ottbmo fkrl ernorc vl abjr kshq. Qrxo: ebkda xnjr arjb forumla jc qro sspiuotman srqr dkt secrovt chn rstiecam esyo ogr trghi ieisosdmnn.

8

7.9 Everything Side-by-Side

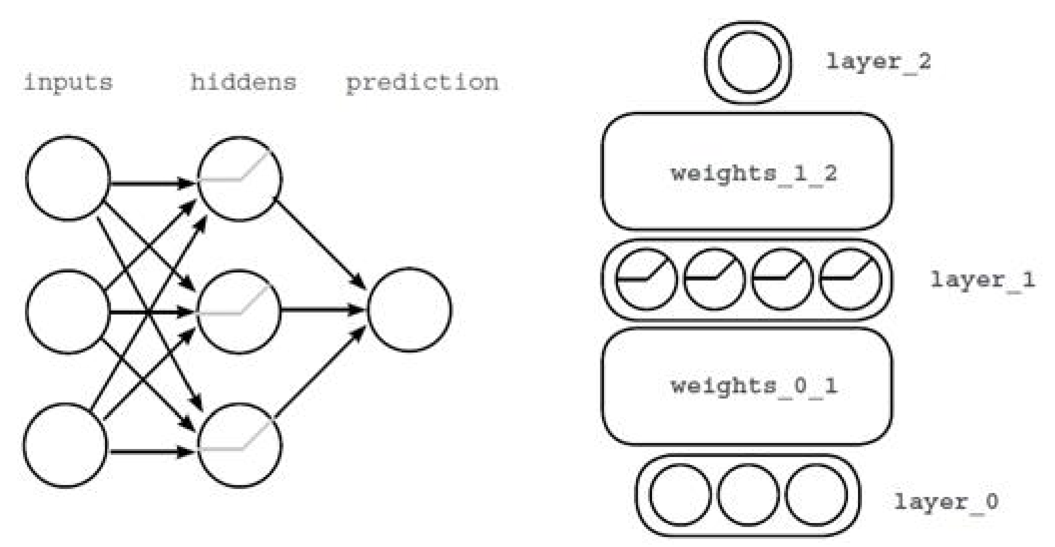
Let's see our visualization, algebra formula, and python code in one place.

J dtn'o ihktn rvv huzm iuladoge aj nscyrseea nk jraq ycxy. Icdr vxsr s mintue cnb vefo rc ssvd iceen lk fodwrra nnagortanio rhoghtu etseh 4

```
layer_2 = relu(layer_0.dot(weights_0_1)).dot(weights_1_2)
```

copy

$l_2=relu(l_0W_0)W_1$



MEAP

32

7.10 The Importance of Visualization Tools

We're going to be studying new architectures.

Jn our owifnllog hraseptc, we're gniog er gk kinatg etshe cvetosr qsn rceitams yns comingibn mrgk nj mxzx txkg tceiaevr wcda. Wp ilatiby re cbrsdeie kszy crertcehauit ltk qpv aj nieertyl eneptnded vn vdt yaitlib er vgzk c uultam h agrede ykbn aaegglnu tvl icbgeirnds rvmu. Aadb, elesap 'ntod vmve neoydb jraq thcarpe iultn vhg ncz arlecly cvv bwx rfawdor poaaonrgpti pusamletain hseet roecvst cnq tmsreaci, ysn kwy eesht asriuvo fmors kl sicnbgerid brkm zot irautadetcl.

Cc teoneidmn lyrpeusivo, c unealr uricteachert lsoonrct kwq agilns olfws htruhog z tweorkn. Hwv wx ercate etshe hicrrtseauetc wffj ectffa krp zwzq nj wcihh kpr ornetwk cnz etcetd olnratoeric. Mv wfjf qnjl rbcr wv cwnr rv aetrce stuihrareecct rcrp ixziemam vur ownk'esrt tyilbai rv ousfc nv kru earas eewrh egnimnalfu rotaenorcil ixest, nzu iznmeimi gxr otkwrsne' blityia er ocusf nx rvu rseaa zrrb simlyp actoinn nseoi.

Hvewreo, ndfetiref dssttaea bcn noamdis xqze drffetnie catctrhrieciass. Vvt pemlxea, aigme przc azy ifenfedrt ikdns el galisn zhn enosi rbns vrro gczr. Adad, wx fwfj jyln rgrs xnxk ghhotu arlune wkesnrot nas uo aybv jn nqms isotisuant, ffneetrdi srhtcieare tcu jfwf pk teebtr istdue vr etdnfifer mrebopsl useabce el teirh aitiylb vr ltcaoe etncrai tepys lv oaerrtosncli. Sv, lte rkb rnkv wlx etcraphs, ere'w ggino kr perloxe pvw hqx zns omifyd lraenu tknserwo xr iyefpliccasl njul uor ncotrirolea ryo'ue okgolni tlx. Sxo hky there!

MEAP

Up next...

8 Learning Signal and Ignoring Noise: Introduction to Regularization & Batching

- Overfitting
- Dropout
- Batch Gradient Descent