Thekeyideaistorandomlydropunits(alongwiththeirconnections)fromtheneural networkduringtraining.Thispreventsunitsfromco-adaptingtoomuch.Duringtraining dropoutsamplesfromanexponentialnumberoft\thinned"networks.Attesttime itiseasytoapproximatetheofaveragingthepredictionsofallthesethinnednetworks bysimplyusingasingleunthinnednetworkthathassmallerweights.Thistly reducesovandgivesmajorimprovementsoverotherregularizationmethods.We showthatdropoutimprovestheperformanceofneuralnetworksonsupervisedlearning tasksinvisionspeechrecognitiondocumentandcomputationalbiology obtainingstate-of-the-artresultsonmanybenchmarkdatasets. Keywords: neuralnetworksregularizationmodelcombinationdeeplearning

1.Introduction

Deepneuralnetworkscontainmultiplenon-linearhiddenlayersandthismakesthemvery expressivemodelsthatcanlearnverycomplicatedrelationshipsbetweentheirinputsand outputs.Withlimitedtrainingdatahowevermanyofthesecomplicatedrelationships willbetheresultofsamplingnoisesotheywillexistinthetrainingsetbutnotinreal testdataevenifitisdrawnfromthesamedistribution.Thisleadstoovandmany methodshavebeendevelopedforreducingit.Theseincludestoppingthetrainingassoonas performanceonavalidationsetstartstogetworseintroducingweightpenaltiesofvarious kindssuchasL1andL2regularizationandsoftweightsharing(NowlanandHinton1992). SrivastavaHintonl

(a)StandardNeuralNet

(b)Afterapplyingdropout. itsposteriorprobabilitygiventhetrainingdata.Thiscansometimesbeapproxima wellforsimpleorsmallmodels(Xiongetal.2011;SalakhutdinovandMnih2008)butwe wouldliketoapproachtheperformanceoftheBayesiangoldstandardusingconsiderably lesscomputation.Weproposetodothisbyapproximatinganequallyweightedgeometric meanofthepredictionsofanexponentialnumberoflearnedmodelsthatshareparameters. Dropoutisatechni providesawayofapproximatelycombiningexponentiallymanytneuralnetwork architecturestly.Theterm\dropout"referstodroppingoutunits(hiddenand visible)inaneuralnetwork.Bydroppingaunitoutwemeantemporarilyremovingitfrom thenetworkalongwithallitsincomingandoutgoingconnectionsasshowninFigure1. 1930

Dropout
Presentwith
probability
p
w
-
(a)Attrainingtime
Always
present
p
w
-
(b)Attesttime
Figure2:
Left
:Aunitattrainingtimethatispresentwithprobability
p
andisconnectedtounits
inthenextlayerwithweights
w
. Applyingdropouttoaneuralnetworkamountstosamplinga\thinned"networkfrom it.Thethinnednetworkconsistsofalltheunitsthatsurviveddropout(Figure1b).A neuralnetwith

n unitscanbeseenasacollectionof2^n possiblethinnedneuralnetworks. Thesenetworksallshareweightssothatthetotalnumberofparametersisst $O(n^2)$or less.Foreachpresentationofeachtrainingcaseanewthinnednetworkissampledand trained.Sotraininganeuralnetworkwithdropoutcanbeseenastrainingacollectionof2^n thinnednetworkswithextensiveweightsharingwhereeachthinnednetworkgetstrained veryrarelyifatall. Attesttimeitisnotfeasibletoexplicitlyaveragethepredictionsfromexponentially manythinnedmodels.Howeveraverysimpleapproximateaveragingmethodworkswellin practice.Theideaistouseasingleneuralnetattesttimewithoutdropout.Theweights ofthisnetworkarescaled-downversionsofthetrainedweights.Ifaunitisretainedwith probability p duringtrainingtheoutgoingweightsofthatunitaremultipliedby p attest timeasshowninFigure2.Thisensuresthatforanyhiddenunitthe expected output(under thedistributionusedtodropunitsattrainingtime)isthesameastheactualoutputat testtime.Bydoingthisscaling2^n networkswithsharedweightscanbecombinedinto asingleneuralnetworktobeusedattesttime.Wefoundthattraininganetworkwith dropoutandusingthisapproximateaveragingmethodattesttimeleadstotly lowergeneralizationerroronawidevarietyofproblemscomparedtotraining withotherregularizationmethods. Theideaofdropoutisnotlimitedtofeed-forwardneuralnets.Itcanbemore appliedtographicalmodelssuchasBoltzmannMachines.Inthispaperweintroduce thedropoutRestrictedBoltzmannMachinemodelandcompareittostandardRestricted BoltzmannMachines(RBM).OurexperimentsshowthatdropoutRBMsarebetterthan standardRBMsincertainrespects. 3.RelatedWork Dropoutcanbeinterpretedasawayofregularizinganeuralnetworkbyaddingnoiseto itshiddenunits.Theideaofaddingnoisetothestatesofunitshaspreviouslybeenusedin thecontextofDenoisingAutoencoders(DAEs)byVincentetal. (20082010)wherenoise

isaddedtotheinputunitsofanautoencoderandthenetworkistrainedtoreconstructthe noise-freeinput.Ourworkextendsthisideabyshowingthatdropoutcanbeely appliedinthehiddenlayersaswellandthatitcanbeinterpretedasaformofmodel averaging.Wealsoshowthataddingnoiseisnotonlyusefulforunsupervisedfeature learningbutcanalsobeextendedtosupervisedlearningproblems.Infactourmethodcan beappliedtootherneuron-basedarchitecturesforexampleBoltzmannMachines.While 5%noisetypicallyworksbestforDAEswefoundthatourweightscalingprocedureapplied attesttimeenablesustousemuchhighernoiselevels.Droppingout20%oftheinputunits and50%ofthehiddenunitswasoftenfoundtobeoptimal. (2013) alsoexploreddeterministicregularizerscorrespondingtoexponential-familynoise

distributionsincludingdropout(whichtheyrefertoas\blankoutnoise").Howeverthey
applynoisetotheinputsandonlyexploremodelswithnohiddenlayers.WangandManning
(2013)proposedamethodforspeedingupdropoutbymarginalizingdropoutnoise.Chen
etal.(2012)exploredmarginalizationinthecontextofdenoisingautoencoders. Thiscanbeseenasminimizin
Roweis(2006);Dekeletal. (2010)exploredanalternatesettingwherethelossisminimized
whenanadversarygetstopickwhichunitstodrop.Hereinsteadofanoisedistribution
themaximumnumberofunitsthatcanbedroppedisHoweverthisworkalsodoes
notexploremodelswithhiddenunits. 4.ModelDescription
Thissectiondescribesthedropoutneuralnetworkmodel.Consideraneuralnetworkwith
L
hiddenlayers.Let
l
2f
1
;:::;L
g
indexthehiddenlayersofthenetwork.Let
z
(
l
)
denote
thevectorofinputsintolayer
l

y
(
l
)
denotethevectorofoutputsfromlayer
l
(
y
(0)
=
x
is
theinput). W
(
l
)
and
b
(
l
)
aretheweightsandbiasesatlayer
l
.Thefeed-forwardoperation
ofastandardneuralnetwork(Figure3a)canbedescribedas(for
l
2f
0

;:::;L

1

g

and anyhiddenunit i

$$z_i^{(l+1)} = w_i^{(l+1)} y^{l} + b_i^{(l+1)};\quad y_i^{(l+1)} = f(z_i^{(l+1)});$$

where f isanyactivationfunctionforexample $f(x) = 1/(1+\exp($

x)). With dropout the feed-forward operation becomes (Figure 3b)

$$r^{(l)}_j \; \mathrm{Bernoulli}(p);$$

$$e\, y^{(l)} = r^{(l)} \; y^{(l)};$$

$$z^{(l+1)}_i = w^{(l+1)}_i e\, y^{l} + b^{(l+1)}_i;$$

$$y^{(l+1)}_i$$

=
f
(
z
(
l
+1)
i
)
:

(a)Standardnetwork
(b)Dropoutnetwork
Figure3:
Comparisonofthebasicoperationsofastandardanddropoutnetwork. Here

denotesanelement-wiseproduct.Foranylayer
l

r
(
l
)
isavectorofindependent
Bernoullirandomvariableseachofwhichhasprobability
p
ofbeing1.Thisvectoris
sampledandmultipliedelement-wisewiththeoutputsofthatlayer
y
(
l
)
tocreatethe
thinnedoutputs
e
y
(
l
)
.Thethinnedoutputsarethenusedasinputtothenextlayer.This
processisappliedateachlayer.Thisamountstosamplingasub-networkfromalarger
network.Forlearningthederivativesofthelossfunctionarebackpropagatedthroughthe
sub-network.Attesttimetheweightsarescaledas
W
(
l
)
test
=
pW
(

l
)
asshowninFigure2. Oneparticularformofregularizationwasfoundtobeespeciallyusefulfordropout|
constrainingthenormoftheincomingweightvectorateachhiddenunittobeupper
boundedbyaconstant
c
.Inotherwordsif
w
representsthevectorofweightsincident
onanyhiddenunittheneuralnetworkwasoptimizedundertheconstraint
jj
w
jj
2

c
.This
constraintwasimposedduringoptimizationbyprojecting
w
ontothesurfaceofaballof
radius
c
whenever
w
wentoutofit.Thisisalsocalledmax-normregularizationsinceit
impliesthatthemaximumvaluethatthenormofanyweightcantakeis
c
.Theconstant
1934

Dropout
c
isatunablehyperparameterwhichisdeterminedusingavalidationset.Max-norm
regularizationhasbeenpreviouslyusedinthecontextofcollaborative(Srebroand
Shraibman2005).Ittypicallyimprovestheperformanceofstochasticgradientdescent
trainingofdeepneuralnetsevenwhennodropoutisused. 5.2UnsupervisedPretraining
Neuralnetworkscanbepretrainedusingstacksof RBMs(HintonandSalakhutdinov2006)
autoencoders(Vincentetal.2010)orDeepBoltzmannMachines(SalakhutdinovandHin-
ton2009).Pretrainingisanewayofmakinguseofunlabeleddata.Pretraining
followedbywithbackpropagationhasbeenshowntogivetperformance
boostsovergfromrandominitializationsincertaincases. CIFAR-10andCIFAR-100:Tinynaturalimages(K
GoogleStreetView(Netzeretal.2011). 1935

SrivastavaHintonKrizhevskySutskeverandSalakhutdinov

AlternativeSplicingdataset:RNAfeaturesforpredictingalternativegenesplicing
(Xiongetal.2011). DataSetDomainDimensionalityTrainingSetTestSet
MNISTVision784(28

28grayscale)60K10K
SVHNVision3072(32

32color)600K26K

CIFAR-10/100Vision3072(32

32color)60K10K
ImageNet(ILSVRC-2012)Vision65536(256

256color)1.2M150K
TIMITSpeech2520(120-dim21frames)1.1Mframes58Kframes
Reuters-RCV1Text2000200K200K
AlternativeSplicingGenetics10142932733
Table1:
Overviewofthedatasetsusedinthispaper. 6.1.1MNIST
Method
Unit
Type
Architecture
Error
%
StandardNeuralNet(Simardetal.2003)Logistic2layers800units1.60
SVMGaussiankernelNANA1.40
DropoutNNLogistic3layers1024units1.35
DropoutNNReLU3layers1024units1.25
DropoutNN+max-normconstraintReLU3layers1024units1.06
DropoutNN+max-normconstraintReLU3layers2048units1.04
DropoutNN+max-normconstraintReLU2layers4096units1.01
DropoutNN+max-normconstraintReLU2layers8192units0.95
DropoutNN+max-normconstraint(Goodfellow
etal.2013)
Maxout
2layers(5

240)
units
0.94
DBN+ng(HintonandSalakhutdinov2006)Logistic500-500-20001.18
DBM+(SalakhutdinovandHinton2009)Logistic500-500-20000.96
DBN+dropoutingLogistic500-500-20000.92
DBM+dropoutetuningLogistic500-500-2000
0.79
Table2:
ComparisonoftmodelsonMNIST. TheMNISTdatasetconsistsof28

28pixelhandwrittendigitimages.Thetaskis
toclassifytheimagesinto10digitclasses.Table2comparestheperformanceofdropout
withothertechniques.Thebestperformingneuralnetworksforthepermutationinvariant
1936

Dropout
settingthatdonotusedropoutorunsupervisedpretrainingachieveanerrorofabout
1
:
60%(Simardetal.2003).Withdropouttheerrorreducesto1
:
35%.Replacinglogistic

unitswithlinearunits(ReLUs)(Jarrettetal.2009)furtherreducestheerrorto
1
:
25%.Addingmax-normregularizationagainreducesitto1
:
06%.Increasingthesizeof
thenetworkleadstobetterresults.Aneuralnetwith2layersand8192unitsperlayer
getsdownto0.95%error.Notethatthisnetworkhasmorethan65millionparametersand
isbeingtrainedonadatasetofsize60000.Traininganetworkofthissizetogivegood
generalizationerrorisveryhardwithstandardregularizationmethodsandearlystopping. (2013)showedtha
0.94%byreplacingReLUunitswithmaxoutunits.Alldropoutnetsuse
p
=0
:
5forhidden
unitsand
p
=0
:
8forinputunits.MoreexperimentaldetailscanbefoundinAppendixB.1. 6.1.2StreetViewHouseNumbers
TheStreetViewHouseNumbers(SVHN)
DataSet(Netzeretal.2011)consistsof
colorimagesofhousenumberscollectedby
GoogleStreetView.Figure5ashowssomeexamplesofimagesfromthisdataset.The
partofthedatasetthatweuseinourexperimentsconsistsof32

32colorimagesroughly
centeredonadigitinahousenumber.Thetaskistoidentifythatdigit. 1989).Thebestarchitecturethatwefounc
fullyconnectedhiddenlayers.AllhiddenunitswereReLUs.Eachconvolutionallayerwas
1937

Method	Error%
BinaryFeatures(WDCH)(Netzeretal.2011)	36.7
HOG(Netzeretal.2011)	15.0
StackedSparseAutoencoders(Netzeretal.2011)	10.3
KMeans(Netzeretal.2011)	9.4
Multi-stageConvNetwithaveragepooling(Sermanetetal.2012)	9.06
Multi-stageConvNet+L2pooling(Sermanetetal.2012)	5.36
Multi-stageConvNet+L4pooling+padding(Sermanetetal.2012)	4.90
ConvNet+max-pooling	3.95
ConvNet+maxpooling+dropoutinfullyconnectedlayers	3.02
ConvNet+stochasticpooling(ZeilerandFergus2013)	2.80
ConvNet+maxpooling+dropoutinalllayers	2.55
ConvNet+maxout(Goodfellowetal.2013)	2.47
HumanPerformance	2.0

Table3:
ResultsontheStreetViewHouseNumbersdataset. Dropoutwasappliedtoallthelayersofthenetworkwithth
denunitbeing
p
=(0
:

9
;
0
:
75
;
0
:
75
;
0
:
5
;
0
:
5
;
0
:
5)forthetlayersofthenetwork(going
frominputtoconvolutionallayerstofullyconnectedlayers).Max-normregularizationwas
usedforweightsinbothconvolutionalandfullyconnectedlayers.Table3comparesthe
resultsobtainedbytmethods.Wethatconvolutionalnetsoutperformother
methods.Thebestperformingconvolutionalnetsthatdonotusedropoutachieveanerror
rateof3
:
95%.Addingdropoutonlytothefullyconnectedlayersreducestheerrorto3
:
02%. Theadditionalgaininperformanceobtainedbyaddingdropoutintheconvolutional
layers(3
:
02%to2
:
55%)isworthnoting.Onemayhavepresumedthatsincetheconvo-
lutionallayersdon'thavealotofparametersovisnotaproblemandtherefore
dropoutwouldnothavemuchHoweverdropoutinthelowerlayersstillhelpsbe-
causeitprovidesnoisyinputsforthehigherfullyconnectedlayerswhichpreventsthem
fromov
6.1.3CIFAR-10andCIFAR-100
TheCIFAR-10andCIFAR-100datasetsconsistof32

32colorimagesdrawnfrom10
and100categoriesrespectively.Figure5bshowssomeexamplesofimagesfromthisdata
set.Adetaileddescriptionofthedatasetsinputpreprocessingnetworkarchitecturesand
otherexperimentaldetailsisgiveninAppendixB.3.Table4showstheerrorrateobtained
bytmethodsonthesedatasets.WithoutanydataaugmentationSnoeketal. (2012)usedBayesianhyperparam
CIFAR-10.Usingdropoutinthefullyconnectedlayersreducesthatto14.32%andadding
dropoutineverylayerfurtherreducestheerrorto12.61%.Goodfellowetal. (2013)showed
thattheerrorisfurtherreducedto11.68%byreplacingReLUunitswithmaxoutunits.On
CIFAR-100dropoutreducestheerrorfrom43.48%to37.20%whichisahugeimprovement.

(a)StreetViewHouseNumbers(SVHN)
(b)CIFAR-10
Figure5:
Samplesfromimagedatasets.Eachrowcorrespondstoatcategory. MethodCIFAR-10CIFAR-100
ConvNet+maxpooling(handtuned)15.6043.48
ConvNet+stochasticpooling(ZeilerandFergus2013)15.1342.51
ConvNet+maxpooling(Snoeketal.2012)14.98-
ConvNet+maxpooling+dropoutfullyconnectedlayers14.3241.26
ConvNet+maxpooling+dropoutinalllayers12.61
37.20
ConvNet+maxout(Goodfellowetal.2013)
11.68
38.57
Table4:
ErrorratesonCIFAR-10andCIFAR-100. 6.1.4ImageNet
ImageNetisadatasetofover15millionlabeledhigh-resolutionimagesbelongingtoroughly
22000categories.Startingin2010aspartofthePascalVisualObjectChallengeanannual
competitioncalledtheImageNetLarge-ScaleVisualRecognitionChallenge(ILSVRC)has
beenheld.AsubsetofImageNetwithroughly1000imagesineachof1000categoriesis
usedinthischallenge.Sincethenumberofcategoriesisratherlargeitisconventionalto
reporttwoerrorrates:top-1andtop-5wherethetop-5errorrateisthefractionoftest
imagesforwhichthecorrectlabelisnotamongtheelabelsconsideredmostprobableby
themodel.Figure6showssomepredictionsmadebyourmodelonafewtestimages. ILSVRC-2010istheonly
mostofourexperimentswereperformedonthisdataset.Table5comparestheperformance
ofmethods.Convolutionalnetswithdropoutoutperformothermethodsbyalarge
margin.ThearchitectureandimplementationdetailsaredescribedindetailinKrizhevsky
etal.(2012). ModelTop-1Top-5
SparseCoding(Linetal.2010)47.128.2
SIFT+FisherVectors(SanchezandPerronnin2011)45.725.7
ConvNet+dropout(Krizhevskyetal.2012)37.517.0
Table5:
ResultsontheILSVRC-2010testset. Model
Top-1
(val)
Top-5
(val)
Top-5
(test)
SVMonFisherVectorsofDenseSIFTandColorStatistics--27.3
AvgofoverFVsofSIFTLBPGISTandCSIFT--26.2
ConvNet+dropout(Krizhevskyetal.2012)40.718.2-
Avgof5ConvNets+dropout(Krizhevskyetal.2012)38.116.416.4
Table6:
ResultsontheILSVRC-2012validation/testset. MethodPhoneErrorRate%
NN(6layers)(Mohamedetal.2010)23.4
DropoutNN(6layers)21.8
DBN-pretrainedNN(4layers)22.7
DBN-pretrainedNN(6layers)(Mohamedetal.2010)22.4
DBN-pretrainedNN(8layers)(Mohamedetal.2010)20.7
mcRBM-DBN-pretrainedNN(5layers)(Dahletal.2010)20.5
DBN-pretrainedNN(4layers)+dropout
19.7
DBN-pretrainedNN(8layers)+dropout

19.7
Table7:
PhoneerrorrateontheTIMITcoretestset. 6.4ComparisonwithBayesianNeuralNetworks
Dropoutcanbeseenasawayofdoinganequally-weightedaveragingofexponentiallymany
modelswithsharedweights.OntheotherhandBayesianneuralnetworks(Neal1996)are
theproperwayofdoingmodelaveragingoverthespaceofneuralnetworkstructuresand
parameters.IndropouteachmodelisweightedequallywhereasinaBayesianneural
networkeachmodelisweightedtakingintoaccountthepriorandhowwellthemodel
thedatawhichisthemorecorrectapproach.Bayesianneuralnetsareextremelyusefulfor
solvingproblemsindomainswheredataisscarcesuchasmedicaldiagnosisgeneticsdrug
discoveryandothercomputationalbiologyapplications.HoweverBayesianneuralnetsare
slowtotrainandtoscaletoverylargenetworksizes.Besidesitisexpensiveto
getpredictionsfrommanylargenetsattesttime.Ontheotherhanddropoutneuralnets
aremuchfastertotrainanduseattesttime.Inthissectionwereportexperimentsthat
compareBayesianneuralnetswithdropoutneuralnetsonasmalldatasetwhereBayesian
neuralnetworksareknowntoperformwellandobtainstate-of-the-artresults.Theaimis
toanalyzehowmuchdoesdropoutlosecomparedtoBayesianneuralnets. Thedatasetthatweuse(Xiongetal.2
taskistopredicttheoccurrenceofalternativesplicingbasedonRNAfeatures.Alternative
splicingisatcauseofcellulardiversityinmammaliantissues.Predictingthe
1941

SrivastavaHintonKrizhevskySutskeverandSalakhutdinov
MethodCodeQuality(bits)
NeuralNetwork(earlystopping)(Xiongetal.2011)440
RegressionPCA(Xiongetal.2011)463
SVMPCA(Xiongetal.2011)487
NeuralNetworkwithdropout567
BayesianNeuralNetwork(Xiongetal.2011)
623
Table8:
ResultsontheAlternativeSplicingDataSet. occurrenceofalternatesplicingincertaintissuesundertconditio
understandingmanyhumandiseases.GiventheRNAfeaturesthetaskistopredictthe
probabilityofthreesplicingrelatedeventsthatbiologistscareabout.Theevaluationmetric
isCodeQualitywhichisameasureofthenegativeKLdivergencebetweenthetargetand
thepredictedprobabilitydistributions(higherisbetter).AppendixB.6includesadetailed
descriptionofthedatasetandthisperformancemetric. (2011)usedBayesianneuralnetsforthistask.Asexpec
netsperformbetterthandropout.Howeverweseethatdropoutimprovessigntly
upontheperformanceofstandardneuralnetsandoutperformsallothermethods.The
challengeinthisdatasetistopreventovsincethesizeofthetrainingsetissmall. Onewaytopreventovgistoredu
standardtechniquessuchasSVMsorlogisticregressioncanbeused.Howeverwithdropout
wewereabletopreventovwithouttheneedtododimensionalityreduction.The
dropoutnetsareverylarge(1000sofhiddenunits)comparedtoafewtensofunitsinthe
Bayesiannetwork.Thisshowsthatdropouthasastrongregularizing
6.5ComparisonwithStandardRegularizers
Severalregularizationmethodshavebeenproposedforpreventingovttinginneuralnet-
works.TheseincludeL2weightdecay(moregenerallyTikhonovregularization(Tikhonov
1943))lasso(Tibshirani1996)KL-sparsityandmax-normregularization.Dropoutcan
beseenasanotherwayofregularizingneuralnetworks.Inthissectionwecompareddropout
withsomeoftheseregularizationmethodsusingtheMNISTdataset. Thevaluesofthyperparametersassocia
constantstargetsparsitydropoutratemax-normupperbound)wereobtainedusinga
validationset.Wefoundthatdropoutcombinedwithmax-normregularizationgivesthe
lowestgeneralizationerror. 7.1onFeatures
(a)Withoutdropout

(b)Dropoutwith
p
=0
:
5. 7.2onSparsity
(a)Withoutdropout
(b)Dropoutwith
p
=0
:
5. (a)Keeping
n

(b)Keeping
pn

Figure9:
ofchangingdropoutratesonMNIST. Figure9bshowsthetesterrorobtainedasafunctionof
p
.Wenoticethatthemagnitude
oferrorsforsmallvaluesof
p
hasreducedbyalotcomparedtoFigure9a(for
p
=0
:
1itfell
from2.7%to1.7%).Valuesof
p
thatareclosetо0.6seemtoperformbestforthischoice
of
pn
butourusualdefaultvalueof0
:
5isclosetooptimal. Theresultsoftheseexperimentsare
showninFigure10.Thenetworkwasgiven
datasetsofsize1005001K5K10K
and50KchosenrandomlyfromtheMNIST
trainingset.Thesamenetworkarchitec-
ture(784-1024-1024-2048-10)wasusedfor
alldatasets.Dropoutwith
p
=0
:
5wasper-
formedatallthehiddenlayersand
p
=0
:
8
attheinputlayer.Itcanbeobservedthat
forextremelysmalldatasets(100500)
dropoutdoesnotgiveanyimprovements. 1946

Dropout
WeagainusetheMNISTdatasetanddobyaveragingthepredictions
of
k
randomlysampledneuralnetworks.Figure11showsthetesterrorrateobtainedfor
tvaluesof
k
.Thisiscomparedwiththeerrorobtainedusingtheweightscaling
method(shownasahorizontalline).Itcanbeseenthataround
k
=50theMonte-Carlo
methodbecomesasgoodastheapproximatemethod.ThereafterTheMonte-Carlomethod
isslightlybetterthantheapproximatemethodbutwellwithinonestandarddeviationof
it.Thissuggeststhattheweightscalingmethodisafairlygoodapproximationofthetrue
modelaverage. 8.1ModelDescription
ConsideranRBMwithvisibleunits
v
2f
0
;
1
g
D
andhiddenunits
h
2f
0
;
1
g
F
.It
thefollowingprobabilitydistribution
P
(
h
;
v
;

)=
1
Z
(

)
exp(
v
>
W
h
+

$a^\top h + b^\top v)$:

Where $\theta = \{W, a, b\}$ represents the model parameters and Z is the partition function. Dropout RBMs are RBMs augmented with a vector of binary random variables $r \in \{0, 1\}^F$. Each random variable r_j takes the value 1 with probability p independent of others. If r_j takes the value 1 the hidden unit h_j is retained otherwise it is dropped from the model. The joint distribution by a Dropout RBM can be expressed as

$$P(r, h, v; p, \theta) =$$

P
(
r
;
p
)
P
(
h
;
v
j
r
;

)
;
P
(
r
;
p
)=
F
Y
j
=1
p
r
j
(1

p
)
1

r
j
;
P
(
h
;
v
j
r
;

)=
1
Z
0
(

;
r
)
exp(
v
>
W
h
+
a
>
h
+
b
>
v
)
F
Y
j
=1
g
(
h
j
;r
j
)
;
g
(
h
j
;r
j
)=
1
(
r
j
=1)+
1
(
r
j
=0)
1
(
h
j
=0)
:
Z

$0(;r)$ is the normalization constant. $g(h_j;r_j)$ imposes the constraint that if $r_j=0$ h_j must be 0. The distribution over h conditioned on v and r is factorial

$$P(h|r;v)=\prod_{j=1}^{F}P(h_j|r_j;v);$$

$P(h_j=1|r$

j
;
v
)=
1
(
r
j
=1)

b
j
+
X
i
W
ij
v
i
! :
1947

SrivastavaHintonKrizhevskySutskeverandSalakhutdinov
(a)Withoutdropout
(b)Dropoutwith
p
=0
:
5. Thedistributionover
v
conditionedon
h
issameasthatofanRBM
P
(
v
j
h
)=
D
Y
i
=1
P
(
v
i
j
h
)
;
P
(

$$v_i = 1 | h) = \left(a_i + \sum_j W_{ij} h_j \right).$$

Conditionedon r thedistributionover $\{v, h\}$ issameasthedistributionthatanRBM wouldimposeexceptthattheunitsforwhich $r_j = 0$ aredroppedfrom h. Therefore the DropoutRBMmodelcanbeseenasamixtureofexponentiallymanyRBMswithshared weightseachusingatsubsetof h.

Figure12ashowsfeatureslearnedbyabinaryRBMwith256hiddenunits.Figure12b showsfeatureslearnedbyadropoutRBMwiththesamenumberofhiddenunits.Features

(a)Withoutdropout

(b)Dropoutwith $p = 0.5$.

9. Marginalizing Dropout

Dropoutcanbeseenasawayofaddingnoisetothestatesofhiddenunitsinaneural network.Inthissectionweexploretheclassofmodelsthatariseasaresultofmarginalizing thisnoise.Thesemodelscanbeseenasdeterministicversionsofdropout.Incontrastto standard(\Monte-Carlo")dropoutthesemodelsdonotneedrandombitsanditispossible

togetgradientsforthemarginalizedlossfunctions.Inthissectionweexplorethese
models. Deterministicalgorithmshavebeenproposedthattrytolearnmodelsthatarerobustto
featuredeletionattesttime(GlobersonandRoweis2006).Marginalizationinthecontext
ofdenoisingautoencodershasbeenexploredpreviously(Chenetal.2012).Themarginal-
izationofdropoutnoiseinthecontextoflinearregressionwasdiscussedinSrivastava(2013). WangandMann
training.vanderMaatenetal. (2013)investigatedtinputnoisedistributionsand

theregularizersobtainedbymarginalizingthisnoise.Wageretal. (2013)describeshow
dropoutcanbeseenasanadaptiveregularizer. y
2
R
N
beavectorof
targets.Linearregressiontriestoa
w
2
R
D
thatminimizes
jj
y

X
w
jj
2
:
Whentheinput
X
isdroppedoutsuchthatanyinputdimensionisretainedwith
probability
p
theinputcanbeexpressedas
R

X
where
R
2f
0
;
1
g
N

D
isarandommatrix
with
R
ij
Bernoulli(

p
)and

denotesanelement-wiseproduct.Marginalizingthenoise
theobjectivefunctionbecomes
minimize
w
E
R
Bernoulli(p)

jj
y

(
R

X
)
w
jj
2

:
Thisreducesto
minimize
w
jj
y

pX
w
jj
2
+
p
(1

p
)
jj

w
jj
2
;
where=(diag(
X
>
X
))
1
=

2
.Thereforedropoutwithlinearregressionisequivalentin
expectationtoridgeregressionwithaparticularformforThisformofessentially
scalestheweightcostforweight
w
i
bythestandarddeviationofthe
i
thdimensionofthe
data.Ifaparticulardatadimensionvariesalottheregularizertriestosqueezeitsweight
more. For
p
closeto1alltheinputsareretainedandtheregularizationconstantissmall.As
moredropoutisdone(bydecreasing
p
)theregularizationconstantgrowslarger. 9.2LogisticRegressionandDeepNetworks
Forlogisticregressionanddeepneuralnetsitishardtoobtainaclosedformmarginalized
model.HoweverWangandManning(2013)showedthatinthecontextofdropoutapplied
tologisticregressionthecorrespondingmarginalizedmodelcanbetrainedapproximately. ForCIFAR-10w
p
=(0
:
9
;
0
:
75
;
0
:
75
;
0
:
5
;
0
:
5
;
0
:
5)goingfromtheinputlayertothe
top.Thevalueof
fortheGaussiandropoutmodelswassettobe
q
1

p
p
.Resultswere
averagedover10trandomseeds. 10.MultiplicativeGaussianNoise
Dropoutinvolvesmultiplyinghiddenactivationsbyrandomvariables

whichtakethevalue1withprobability p and0otherwise.Thisideacanbegeneralized
bymultiplyingtheactivationswithrandomvariablesdrawnfromotherdistributions.We
recentlydiscoveredthatmultiplyingbyarandomvariabledrawnfrom $N(1;1)$ worksjust
aswellorperhapsbetterthanusingBernoullinoise.Thisnewformofdropoutamounts
toaddingaGaussiandistributedrandomvariablewithzeromeanandstandarddeviation
equaltotheactivationoftheunit.Thatiseachhiddenactivation h_i isperturbedto $h_i + h_i r$ where r $(0;1)$ orequivalently $h_i r^0$ where r^0 $(1;1)$.Wecangeneralize
thisto r^0 $(1\ 2)$ where
becomesanadditionalhyperparametertotunejustlike p
wasinthestandard(Bernoulli)dropout.Theexpectedvalueoftheactivationsremains
unchangedthereforenoweightscalingisrequiredattesttime. FortheGaussianmultiplicativenoiseifweset $2 = (1\ p) = p$ weendupmultiplying

h
i
byarandomvariable
r
g
where
E
[
r
g
]=1and
Var
[
r
g
]=(1

p
)
=p
.Thereforeboth
formsofdropoutcanbesetupsothattherandomvariablebeingmultipliedbyhasthe
samemeanandvariance.Howevergiventheseandsecondordermoments
r
g
hasthe
highestentropyand
r
b
hasthelowest.Boththeseextremesworkwellalthoughpreliminary
experimentalresultsshowninTable10suggestthatthehighentropycasemightwork
slightlybetter.Foreachlayerthevalueof
intheGaussianmodelwassettobe
q
1

p
p
usingthe
p
fromthecorrespondinglayerintheBernoullimodel. Forinputlayersthechoicedependsonthekindofinput.F
patchesorspeechframes)atypicalvalueis0
:
8.Forhiddenlayersthechoiceof
p
iscoupled
withthechoiceofnumberofhiddenunits
n
.Smaller
p
requiresbig
n
whichslowsdown

thetrainingandleadstounderLarge
p
maynotproduceenoughdropouttoprevent
ov
AppendixB.DetailedDescriptionofExperimentsandDataSets
. Thissectiondescribesthenetworkarchitecturesandtrainingdetailsfortheexperimental
resultsreportedinthispaper.Thecodeforreproducingtheseresultscanbeobtainedfrom
http://www.cs.toronto.edu/
~
nitish/dropout
.TheimplementationisGPU-based.We
usedtheexcellentCUDAlibraries|cudamat(Mnih2009)andcuda-convnet(Krizhevsky
etal.2012)toimplementournetworks. 1953

ThearchitecturesshowninFigure4includeallcombinationsof23and4layernetworks
with1024and2048unitsineachlayer.Thustherearesixarchitecturesinall.Forallthe
architectures(includingtheonesreportedinTable2)weused
p
=0
:
5inallhiddenlayers
and
p
=0
:
8intheinputlayer.Amomentumof0
:
95andweightconstraintswith
c
=2
wasusedinallthelayers. B.2SVHN
TheSVHNdatasetconsistsofapproximately600000trainingimagesand26000test
images.Thetrainingsetconsistsoftwoparts|Astandardlabeledtrainingsetandanother
setoflabeledexamplesthatareeasy.Avalidationsetwasconstructedbytakingexamples
fromboththeparts.Two-thirdsofitweretakenfromthestandardset(400perclass)and
one-thirdfromtheextraset(200perclass)atotalof6000samples.Thissameprocess
isusedbySermanetetal. (2012).TheinputswereRGBpixelsnormalizedtohavezero
meanandunitvariance.Otherpreprocessingtechniquessuchasglobalorlocalcontrast
normalizationorZCAwhiteningdidnotgiveanynoticeableimprovements. Eachconvolutionallayerhasa5

5receptiveappliedwithastrideof1pixel.Each
maxpoolinglayerpools3

3regionsatstridesof2pixels.Theconvolutionallayersare
followedbytwofullyconnectedhiddenlayershaving2048unitseach.Allunitsusethe
linearactivationfunction.Dropoutwasappliedtoallthelayersofthenetwork
withtheprobabilityofretainingtheunitbeing
p
=(0
:
9
;

0
:
75
;
0
:
75
;
0
:
5
;
0
:
5
;
0
:
5)forthe
tlayersofthenetwork(goingfrominputtoconvolutionallayerstofullyconnected
layers).Inadditionthemax-normconstraintwith
c
=4wasusedforalltheweights.A
momentumof0
:
95wasusedinallthelayers.Thesehyperparametersweretunedusinga
validationset.Sincethetrainingsetwasquitelargewedidnotcombinethevalidation
setwiththetrainingsetfortraining.Wereportedtesterrorofthemodelthathad
smallestvalidationerror. B.4TIMIT
TheopensourceKalditoolkit(Poveyetal.2011)wasusedtopreprocessthedataintolog-
banks.Amonophonesystemwastrainedtodoaforcedalignmentandtogetlabelsfor
speechframes.Dropoutneuralnetworksweretrainedonwindowsof21consecutiveframes
topredictthelabelofthecentralframe.Nospeakerdependentoperationswereperformed. Max-normconstra
c
=4wasusedinallthelayers.Amomentumof0.95witha
highlearningrateof0.1wasused.Thelearningratewasdecayedas

0
(1+
t=T
)

1
.For
DBNpretrainingwetrainedRBMsusingCD-1.Thevarianceofeachinputunitforthe
GaussianRBMwasto1.ForingtheDBNwithdropoutwefoundthatin
ordertogetthebestresultsitwasimportanttouseasmallerlearningrate(about0.01). B.5Reuters
TheReutersRCV1corpuscontainsmorethan800000documentscategorizedinto103
classes.Theseclassesarearrangedinatreehierarchy.Wecreatedasubsetofthisdataset
consistingof402738articlesandavocabularyof2000wordscomprisingof50categories
inwhicheachdocumentbelongstoexactlyoneclass.Thedatawassplitintoequalsized
trainingandtestsets.Wetriedmanynetworkarchitecturesandfoundthatdropoutgave
improvementsinaccuracyoverallofthem.Howevertheimprovementwas

notastasthatfortheimageandspeechdatasets.Thismightbeexplainedby thefactthatthisdatasetisquitebig(morethan200000trainingexamples)andov isnotaveryseriousproblem.

B.6AlternativeSplicing

Thealternativesplicingdatasetconsistsofdatafor3665cassetteexons1014RNAfeatures and4tissuetypesderivedfrom27mousetissues.Foreachinputthetargetconsistsof4 softmaxunits(onefortissuetype).Eachsoftmaxunithas3states(inc, exc, nc)whichare ofthebiologicalimportance.Foreachsoftmaxunittheaimistopredictadistributionover these3statesthatmatchestheobserveddistributionfromwetlabexperimentsasclosely aspossible.TheevaluationmetricisCodeQualitywhichisas

$$\sum_{i=1}^{|\text{datapoints}|}\sum_{t\in\text{tissuetypes}}\sum_{s\in\{inc,exc,nc\}} p^{s}_{i,t}\log\left(\frac{q^{s}_{t}(r_i)}{p^{s}}\right);$$

where $p^{s}_{i,t}$ isthetargetprobabilityforstate s andtissuetype

t
ininput
i
;
q
s
t
(
r
i
)isthe
predictedprobabilityforstate
s
intissuetype
t
forinput
r
i
and
p
s
istheaverageof
p
s
i;t
over
i
and
t
. Avalueof
p
=0
:
5wasusedforthehiddenlayerand
p
=0
:
7fortheinputlayer.Max-norm
regularizationwithhighdecayinglearningrateswasused.Resultswereaveragedacrossthe
same5foldsusedbyXiongetal.(2011). MachineLearning
81(2):149{1782010. K.JarrettK.KavukcuogluM.RanzatoandY.LeCun.Whatisthebestmulti-stage
architectureforobjectrecognition?In
ProceedingsoftheInternationalConferenceon
ComputerVision(ICCV'09)
.IEEE2009. NeuralComputa-
tion
1(4):541{5511989. ProceedingsoftheNationalAcademyofSciences
107(4):1452{14572010. P.SermanetS.ChintalaandY.LeCun.Convolutionalneuralnetworksappliedtohe
numbersdigitIn
InternationalConferenceonPatternRecognition(ICPR
2012)
2012. JournaloftheRoyal
StatisticalSociety.SeriesB.Methodological

58(1):267{2881996. Bioinformatics
27(18):2554{25622011.