

DATA ANALYTICS USING DEEP LEARNING GT 8803 // Fall 2019 // Joy Arulraj

LECTURE #13:TRAINING NEURAL NETWORKS (PT 2)

CREATING THE NEXT®

ADMINISTRIVIA

- Reminders
 - Assignment 1 grades released
 - Project progress reports due in two weeks
 - Assignment 2 due in three weeks



LAST TIME: ACTIVATION FUNCTIONS



Leaky ReLU $\max(0.1x, x)$



 $\begin{array}{l} \textbf{Maxout} \\ \max(w_1^T x + b_1, w_2^T x + b_2) \end{array}$



Georgia Tech

LAST TIME: ACTIVATION FUNCTIONS



Leaky ReLU $\max(0.1x, x)$



 $\begin{array}{l} \textbf{Maxout} \\ \max(w_1^T x + b_1, w_2^T x + b_2) \end{array}$





LAST TIME: WEIGHT INITIALIZATION



Initialization too small:

Activations go to zero, gradients also zero, No learning =(

Initialization too big: Activations saturate (for tanh), Gradients zero, no learning =(



Nice distribution of activations at all layers, Learning proceeds nicely



LAST TIME: DATA PREPROCESSING





LAST TIME: BATCH NORMALIZATION

[loffe and Szegedy, 2015]

Input:
$$x: N \times D$$

Learnable scale and shift parameters:

 $\gamma, eta: D$

Learning $\gamma = \sigma$ $\beta = \mu$ will recover the identity function!





TODAY'S AGENDA

- Improve your training error:
 - Optimizers
 - Learning rate schedules
- Improve your test error
 - Regularization
 - Choosing hyperparameters





OPTIMIZATION



OPTIMIZATION: SGD

```
# Vanilla Gradient Descent
while True:
   weights_grad = evaluate_gradient(loss_fun, data, weights)
   weights += - step_size * weights_grad # perform parameter update
```





What if loss changes quickly in one direction and slowly in another? What does gradient descent do?



Loss function has high **condition number**: ratio of largest to smallest singular value of the Hessian matrix is large



What if loss changes quickly in one direction and slowly in another? What does gradient descent do?

Very slow progress along shallow dimension, jitter along steep direction



Loss function has high **condition number**: ratio of largest to smallest singular value of the Hessian matrix is large



What if the loss function has a **local minima** or **saddle point**?





What if the loss function has a **local minima** or **saddle point**?

Zero gradient, gradient descent gets stuck





What if the loss function has a **local minima** or **saddle point**?

Saddle points much more common than local minima in high dimension

Dauphin et al, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization", NIPS 2014





Our gradients come from minibatches so they can be noisy!

$$L(W) = \frac{1}{N} \sum_{i=1}^{N} L_i(x_i, y_i, W)$$

$$\nabla_W L(W) = \frac{1}{N} \sum_{i=1}^N \nabla_W L_i(x_i, y_i, W)$$





SGD

 $x_{t+1} = x_t - \alpha \nabla f(x_t)$

while True: dx = compute_gradient(x) x -= learning_rate * dx SGD+Momentum

 $v_{t+1} = \rho v_t + \nabla f(x_t)$ $x_{t+1} = x_t - \alpha v_{t+1}$ vx = 0
while True:
dx = compute_gradient(x)
vx = rho * vx + dx
x -= learning_rate * vx

- Build up "velocity" as a running mean of gradients
- Rho gives "friction"; typically rho=0.9 or 0.99

Sutskever et al, "On the importance of initialization and momentum in deep learning", ICML 2013



SGD+Momentum

$v_{t+1} =$	$= ho v_t - lpha abla j$	$f(x_t)$
$x_{t+1} =$	$x_t + v_{t+1}$	

vx = 0
while True:
 dx = compute_gradient(x)
 vx = rho * vx - learning_rate * dx
 x += vx

SGD+Momentum

```
v_{t+1} = \rho v_t + \nabla f(x_t)
x_{t+1} = x_t - \alpha v_{t+1}
vx = 0
while True:
dx = compute_gradient(x)
vx = rho * vx + dx
x -= learning_rate * vx
```

You may see SGD+Momentum formulated different ways, but they are equivalent - given same sequence of x

Sutskever et al, "On the importance of initialization and momentum in deep learning", ICML 2013



Gradient Noise





Momentum update:



Gradient

Combine gradient at current point with velocity to get step used to update weights



Nesterov, "A method of solving a convex programming problem with convergence rate O(1/k^2)", 1983 Nesterov, "Introductory lectures on convex optimization: a basic course", 2004 Sutskever et al, "On the importance of initialization and momentum in deep learning", ICML 2013



Momentum update:



Velocity actual step

Nesterov Momentum

Gradient

Combine gradient at current point with velocity to get step used to update weights

"Look ahead" to the point where updating using velocity would take us; compute gradient there and mix it with velocity to

Nesterov, "A method of solving a convex programming problem with convergence rate O(1/k^2)", 1983 Nesterov, "Introductory lectures on convex optimization: a basic course", 2004 Sutskever et al, "On the importance of initialization and momentum in deep learning", ICML 2013

⁻ 8803 // FALL 2018



$$v_{t+1} = \rho v_t - \alpha \nabla f(x_t + \rho v_t)$$
$$x_{t+1} = x_t + v_{t+1}$$



"Look ahead" to the point where updating using velocity would take us; compute gradient there and mix it with velocity to get actual update direction



$$v_{t+1} = \rho v_t - \alpha \nabla f(x_t + \rho v_t)$$

$$x_{t+1} = x_t + v_{t+1}$$
Annoying, usually we want update in terms of $x_t, \nabla f(x_t)$
Gradient Velocity actual step

"Look ahead" to the point where updating using velocity would take us; compute gradient there and mix it with velocity to get actual update direction



$$\begin{aligned} v_{t+1} &= \rho v_t - \alpha \nabla f(x_t + \rho v_t) \\ x_{t+1} &= x_t + v_{t+1} \end{aligned}$$
Change of variables $\tilde{x}_t &= x_t + \rho v_t$ and rearrange:
 $v_{t+1} &= \rho v_t - \alpha \nabla f(\tilde{x}_t) \\ \tilde{x}_{t+1} &= \tilde{x}_t - \rho v_t + (1 + \rho) v_{t+1} \\ &= \tilde{x}_t + v_{t+1} + \rho(v_{t+1} - v_t) \end{aligned}$
Annoying, usually we want update in terms of $x_t, \nabla f(x_t)$
Gradient Velocity actual step $v_{t+1} = \tilde{x}_t - \rho v_t + (1 + \rho) v_{t+1} \\ &= \tilde{x}_t + v_{t+1} + \rho(v_{t+1} - v_t) \end{aligned}$
"Look ahead" to the point where updating using velocity would take us; compute gradient there and mix it with velocity to get actual update direction



$$v_{t+1} = \rho v_t - \alpha \nabla f(x_t + \rho v_t)$$
$$x_{t+1} = x_t + v_{t+1}$$

Annoying, usually we want update in terms of $x_t, \nabla f(x_t)$

Change of variables
$$\, \tilde{x}_t = x_t + \rho v_t \,$$
 and rearrange:

$$v_{t+1} = \rho v_t - \alpha \nabla f(\tilde{x}_t)$$

$$\tilde{x}_{t+1} = \tilde{x}_t - \rho v_t + (1+\rho)v_{t+1}$$

$$= \tilde{x}_t + v_{t+1} + \rho(v_{t+1} - v_t)$$







```
grad_squared = 0
while True:
    dx = compute_gradient(x)
    grad_squared += dx * dx
    x -= learning_rate * dx / (np.sqrt(grad_squared) + 1e-7)
```

Added element-wise scaling of the gradient based on the historical sum of squares in each dimension

"Per-parameter learning rates" or "adaptive learning rates"

Duchi et al, "Adaptive subgradient methods for online learning and stochastic optimization", JMLR 2011





Q: What happens with AdaGrad?







Q: What happens with AdaGrad?

Progress along "steep" directions is damped; progress along "flat" directions is accelerated





Q2: What happens to the step size over long time?



RMSPROP: "LEAKY ADAGRAD"



Tieleman and Hinton, 2012



RMSPROP





ADAM (ALMOST)

```
first_moment = 0
second_moment = 0
while True:
    dx = compute_gradient(x)
    first_moment = beta1 * first_moment + (1 - beta1) * dx
    second_moment = beta2 * second_moment + (1 - beta2) * dx * dx
    x -= learning_rate * first_moment / (np.sqrt(second_moment) + 1e-7))
```



ADAM (ALMOST)



Sort of like RMSProp with momentum

Q: What happens at first timestep?



ADAM (FULL FORM)



Bias correction for the fact that first and second moment estimates start at zero



ADAM (FULL FORM)



Bias correction for the fact that first and second moment estimates start at zero

Adam with beta1 = 0.9, beta2 = 0.999, and learning_rate = 1e-3 or 5e-4 is a great starting point for many models!








GT 8803 // FALL 2018



LEARNING RATE SCHEDULES



GT 8803 // FALL 2018

SGD, SGD+Momentum, Adagrad, RMSProp, Adam all have **learning rate** as a hyperparameter.



Q: Which one of these learning rates is best to use?



SGD, SGD+Momentum, Adagrad, RMSProp, Adam all have **learning rate** as a hyperparameter.



Q: Which one of these learning rates is best to use?

A: All of them! Start with large learning rate and decay over time





Step: Reduce learning rate at a few fixed points. E.g. for ResNets, multiply LR by 0.1 after epochs 30, 60, and 90.





Loshchilov and Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts", ICLR 2017 Radford et al, "Improving Language Understanding by Generative Pre-Training", 2018 Feichtenhofer et al, "SlowFast Networks for Video Recognition", arXiv 2018 Child at al, "Generating Long Sequences with Sparse Transformers", arXiv 2019

Step: Reduce learning rate at a few fixed points. E.g. for ResNets, multiply LR by 0.1 after epochs 30, 60, and 90.

Cosine:
$$\alpha_t = \frac{1}{2} \alpha_0 \left(1 + \cos(t\pi/T) \right)$$

 $\begin{array}{l} \alpha_0 \\ \alpha_t \\ \vdots \text{Learning rate at epoch t} \\ T \\ \vdots \text{Total number of epochs} \end{array}$





Loshchilov and Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts", ICLR 2017 Radford et al, "Improving Language Understanding by Generative Pre-Training", 2018 Feichtenhofer et al, "SlowFast Networks for Video Recognition", arXiv 2018 Child at al, "Generating Long Sequences with Sparse Transformers", arXiv 2019

Step: Reduce learning rate at a few fixed points. E.g. for ResNets, multiply LR by 0.1 after epochs 30, 60, and 90.

Cosine:
$$\alpha_t = \frac{1}{2} \alpha_0 \left(1 + \cos(t\pi/T) \right)$$

 $\begin{array}{l} \alpha_0 \\ \alpha_t \\ T \end{array} : \text{Learning rate at epoch t} \\ T \end{array}$





Devlin et al, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", 2018

Step: Reduce learning rate at a few fixed points. E.g. for ResNets, multiply LR by 0.1 after epochs 30, 60, and 90.

Cosine:
$$\alpha_t = \frac{1}{2} \alpha_0 \left(1 + \cos(t\pi/T)\right)$$

Linear: $\alpha_t = \alpha_0 (1 - t/T)$

 $lpha_0$: Initial learning rate $lpha_t$: Learning rate at epoch t T : Total number of epochs





Step: Reduce learning rate at a few fixed points. E.g. for ResNets, multiply LR by 0.1 after epochs 30, 60, and 90.

Cosine:
$$\alpha_t = \frac{1}{2} \alpha_0 \left(1 + \cos(t\pi/T)\right)$$

Linear: $\alpha_t = \alpha_0 (1 - t/T)$

 $\alpha_t = \alpha_0 / \sqrt{t}$

 $lpha_0$: Initial learning rate $lpha_t$: Learning rate at epoch t T : Total number of epochs

Vaswani et al, "Attention is all you need", NIPS 2017



LEARNING RATE DECAY: LINEAR WARMUP



High initial learning rates can make loss explode; linearly increasing learning rate from 0 over the first ~5000 iterations can prevent this

Empirical rule of thumb: If you increase the batch size by N, also scale the initial learning rate by N

Goyal et al, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour", arXiv 2017



FIRST-ORDER OPTIMIZATION





FIRST-ORDER OPTIMIZATION









second-order Taylor expansion:

$$J(\boldsymbol{\theta}) \approx J(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_0) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

Solving for the critical point we obtain the Newton parameter update:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_0 - \boldsymbol{H}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_0)$$

Q: Why is this bad for deep learning?



second-order Taylor expansion:

$$J(\boldsymbol{\theta}) \approx J(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_0) + \frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

Solving for the critical point we obtain the Newton parameter update:

$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_0 - \boldsymbol{H}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_0)$$

Hessian has O(N^2) elements Inverting takes O(N^3) N = (Tens or Hundreds of) Millions

Q: Why is this bad for deep learning?



$$\boldsymbol{\theta}^* = \boldsymbol{\theta}_0 - \boldsymbol{H}^{-1} \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}_0)$$

Quasi-Newton methods (**BGFS** most popular): instead of inverting the Hessian ($O(n^3)$), approximate inverse Hessian with rank 1 updates over time ($O(n^2)$ each).

L-BFGS (Limited memory BFGS): Does not form/store the full inverse Hessian.





Usually works very well in full batch, deterministic mode i.e. if you have a single, deterministic f(x) then L-BFGS will probably work very nicely

Does not transfer very well to mini-batch setting. Gives bad results. Adapting second-order methods to large-scale, stochastic setting is an active area of research.

Le et al, "On optimization methods for deep learning, ICML 2011" Ba et al, "Distributed second-order optimization using Kronecker-factored approximations", ICLR 2017



IN PRACTICE

- Adam is a good default choice in many cases; it often works ok even with constant learning rate
- SGD+Momentum can outperform Adam but may require more tuning of LR and schedule Try cosine schedule, very few hyperparameters!
- If you can afford to do full batch updates then try out L-BFGS (and don't forget to disable all sources of noise)





TEST ERROR



GT 8803 // FALL 2018

BEYOND TRAINING ERROR



Better optimization algorithms help reduce training loss

But we really care about error on new data - how to reduce the gap?



EARLY STOPPING: ALWAYS DO THIS



Stop training the model when accuracy on the validation set decreases Or train for a long time, but always keep track of the model snapshot that worked best on val



MODEL ENSEMBLES

- 1. Train multiple independent models
- 2. At test time average their results (Take average of predicted probability distributions, then choose

argmax)



MODEL ENSEMBLES: TIPS AND TRICKS

Instead of training independent models, use multiple snapshots of a single model during training!



Loshchilov and Hutter, "SGDR: Stochastic gradient descent with restarts", arXiv 2016 Huang et al, "Snapshot ensembles: train 1, get M for free", ICLR 2017 Figures copyright Yixuan Li and Geoff Pleiss, 2017. Reproduced with permission.



GT 8803 // FALL 2018

MODEL ENSEMBLES: TIPS AND TRICKS

Instead of training independent models, use multiple snapshots of a single model during training!



Loshchilov and Hutter, "SGDR: Stochastic gradient descent with restarts", arXiv 2016 Huang et al, "Snapshot ensembles: train 1, get M for free", ICLR 2017 Figures copyright Yixuan Li and Geoff Pleiss, 2017. Reproduced with permission. Cyclic learning rate schedules can make this work even better!



300

MODEL ENSEMBLES: TIPS AND TRICKS

Instead of using actual parameter vector, keep a moving average of the parameter vector and use that at test time (Polyak averaging)

```
while True:
    data_batch = dataset.sample_data_batch()
    loss = network.forward(data_batch)
    dx = network.backward()
    x += - learning_rate * dx
    x_test = 0.995*x_test + 0.005*x # use_for_test_set
```

Polyak and Juditsky, "Acceleration of stochastic approximation by averaging", SIAM Journal on Control and Optimization, 1992.



HOW TO IMPROVE SINGLE-MODEL PERFORMANCE?



Regularization





REGULARIZATION



GT 8803 // FALL 2018

REGULARIZATION: ADD TERM TO LOSS

$$L = rac{1}{N} \sum_{i=1}^N \sum_{j
eq y_i} \max(0, f(x_i; W)_j - f(x_i; W)_{y_i} + 1) + \lambda R(W)$$

In common use: L2 regularization L1 regularization Elastic net (L1 + L2)

$$egin{aligned} R(W) &= \sum_k \sum_l W_{k,l}^2 & ext{(Weight decay)} \ R(W) &= \sum_k \sum_l |W_{k,l}| \ R(W) &= \sum_k \sum_l eta W_{k,l}^2 + |W_{k,l}| \end{aligned}$$



In each forward pass, randomly set some neurons to zero Probability of dropping is a hyperparameter; 0.5 is common





Srivastava et al, "Dropout: A simple way to prevent neural networks from overfitting", JMLR 2014



p = 0.5 # probability of keeping a unit active. higher = less dropout

```
def train_step(X):
    """ X contains the data """
```

```
# forward pass for example 3-layer neural network
H1 = np.maximum(0, np.dot(W1, X) + b1)
U1 = np.random.rand(*H1.shape)
```

backward pass: compute gradients... (not shown)
perform parameter update... (not shown)

Example forward pass with a 3-layer network using dropout





How can this possibly be a good idea?



Forces the network to have a redundant representation; Prevents co-adaptation of features





How can this possibly be a good idea?



Another interpretation:

Dropout is training a large **ensemble** of models (that share parameters).

Each binary mask is one model

An FC layer with 4096 units has 2⁴⁰⁹⁶ ~ 10¹²³³ possible masks! Only ~ 10⁸² atoms in the universe...



Dropout makes our output random!



Want to "average out" the randomness at test-time

$$y = f(x) = E_z [f(x, z)] = \int p(z) f(x, z) dz$$

But this integral seems hard ...



Want to approximate the integral

$$y = f(x) = E_z [f(x, z)] = \int p(z)f(x, z)dz$$

Consider a single neuron.





Want to approximate the integral

$$y = f(x) = E_z [f(x, z)] = \int p(z)f(x, z)dz$$

Consider a single neuron.

At test time we have:

$$E[a] = w_1 x + w_2 y$$



а

Want to approximate the integral

 W_2

$$y = f(x) = E_z [f(x, z)] = \int p(z)f(x, z)dz$$

Consider a single neuron.

At test time we have:

During training we have:

$$E[a] = w_1 x + w_2 y$$

$$E[a] = \frac{1}{4}(w_1 x + w_2 y) + \frac{1}{4}(w_1 x + 0y)$$

$$+ \frac{1}{4}(0x + 0y) + \frac{1}{4}(0x + w_2 y)$$

$$= \frac{1}{2}(w_1 x + w_2 y)$$



 W_1

Х
DROPOUT: TEST TIME

Want to approximate the integral

$$y = f(x) = E_z [f(x, z)] = \int p(z)f(x, z)dz$$



Consider a single neuron.

At test time we have:

During training we have:

At test time, **multiply** by dropout probability

$$E[a] = w_1 x + w_2 y$$

$$E[a] = \frac{1}{4}(w_1 x + w_2 y) + \frac{1}{4}(w_1 x + 0y)$$

$$+ \frac{1}{4}(0x + 0y) + \frac{1}{4}(0x + w_2 y)$$

$$= \frac{1}{2}(w_1 x + w_2 y)$$



DROPOUT: TEST TIME

def predict(X):

```
# ensembled forward pass
H1 = np.maximum(0, np.dot(W1, X) + b1) * p # NOTE: scale the activations
H2 = np.maximum(0, np.dot(W2, H1) + b2) * p # NOTE: scale the activations
out = np.dot(W3, H2) + b3
```

At test time all neurons are active always => We must scale the activations so that for each neuron: <u>output at test time</u> = <u>expected output at training time</u>



""" Vanilla Dropout: Not recommended implementation (see notes below) """

p = 0.5 # probability of keeping a unit active. higher = less dropout

def train step(X): """ X contains the data """ # forward pass for example 3-layer neural network H1 = np.maximum(0, np.dot(W1, X) + b1)U1 = np.random.rand(*H1.shape) H1 *= U1 # drop! drop in forward pass H2 = np.maximum(0, np.dot(W2, H1) + b2)U2 = np.random.rand(*H2.shape) < p # second dropout mask H2 *= U2 # drop! out = np.dot(W3, H2) + b3 # backward pass: compute gradients... (not shown) # perform parameter update... (not shown) def predict(X): # ensembled forward pass scale at test time H1 = np.maximum(0, np.dot(W1, X) + b1) * p # NOTE: scale the activations H2 = np.maximum(0, np.dot(W2, H1) + b2) * p # NOTE: scale the activations out = np.dot(W3, H2) + b3



GT 8803 // FALL 2018

DROPOUT: SUMMARY

MORE COMMON: "INVERTED DROPOUT"

p = 0.5 # probability of keeping a unit active. higher = less dropout





REGULARIZATION: A COMMON PATTERN

Training: Add some kind of randomness

$$y = f_W(x, z)$$

Testing: Average out randomness (sometimes approximate)

$$y = f(x) = E_z [f(x, z)] = \int p(z) f(x, z) dz$$



REGULARIZATION: A COMMON PATTERN

Training: Add some kind of randomness

$$y = f_W(x, z)$$

Testing: Average out randomness (sometimes approximate)

$$y = f(x) = E_z[f(x,z)] = \int p(z)f(x,z)dz$$

Example: Batch Normalization

Training: Normalize using stats from random minibatches

Testing: Use fixed global stats to normalize



REGULARIZATION: DATA AUGMENTATION





REGULARIZATION: DATA AUGMENTATION



Transform image



DATA AUGMENTATION Horizontal Flips







Random crops and scales Training: sample random crops / scales ResNet:

- 1. Pick random L in range [256, 480]
- 2. Resize training image, short side = L
- 3. Sample random 224 x 224 patch





Random crops and scales Training: sample random crops / scales ResNet:

- 1. Pick random L in range [256, 480]
- 2. Resize training image, short side = L
- 3. Sample random 224 x 224 patch



ResNet:

- 1. Resize image at 5 scales: {224, 256, 384, 480, 640}
- 2. For each size, use 10 224 x 224 crops: 4 corners + center, + flips





Color Jitter

Simple: Randomize contrast and brightness







Color Jitter

Simple: Randomize contrast and brightness





More Complex:

- 1. Apply PCA to all [R, G, B] pixels in training set
- Sample a "color offset" along principal component directions
- 3. Add offset to all pixels of a training image

([Krizhevsky et al. 2012], ResNet, etc)



Get creative for your problem!

Random mix/combinations of :

- translation
- rotation
- stretching
- shearing,
- lens distortions, ... (go crazy)



REGULARIZATION: A COMMON PATTERN

Training: Add random noise **Testing**: Marginalize over the noise

Examples:

Dropout Batch Normalization Data Augmentation



REGULARIZATION: DROPCONNECT

Training: Drop connections between neurons (set weights to 0) **Testing**: Use all the connections

Examples:

Dropout Batch Normalization Data Augmentation DropConnect





Wan et al, "Regularization of Neural Networks using DropConnect", ICML 2013



REGULARIZATION: FRACTIONAL MAX POOLING

Training: Use randomized pooling regions **Testing**: Average predictions from several regions

Examples:

Dropout Batch Normalization Data Augmentation DropConnect Fractional Max Pooling



Graham, "Fractional Max Pooling", arXiv 2014



REGULARIZATION: STOCHASTIC DEPTH

Training: Skip some layers in the network **Testing**: Use all the layers

Examples:

Dropout Batch Normalization Data Augmentation DropConnect Fractional Max Pooling Stochastic Depth

Huang et al, "Deep Networks with Stochastic Depth", ECCV 2016



REGULARIZATION: CUTOUT

Training: Train on random blends of images Testing: Use full image

Examples:

Tech

Dropout Batch Normalization Data Augmentation DropConnect Fractional Max Pooling Stochastic Depth Cutout





Works very well for small datasets like CIFAR, less common for large datasets like ImageNet

REGULARIZATION: MIXUP

Training: Set random image regions to zero **Testing**: Use original images

Examples:

Dropout Batch Normalization Data Augmentation DropConnect Fractional Max Pooling Stochastic Depth Cutout Mixup







Target label: cat: 0.4 dog: 0.6

CNN

Randomly blend the pixels of pairs of training images, e.g. 40% cat, 60% dog

Zhang et al, "mixup: Beyond Empirical Risk Minimization", ICLR 2018



REGULARIZATION: SUMMARY

Training: Add random noise **Testing**: Marginalize over the noise

Examples:

Dropout

Batch Normalization

Data Augmentation

DropConnect Fractional Max Pooling Stochastic Depth

Cutout

Mixup

- Consider dropout for large fully-connected layers
- Batch normalization and data augmentation almost always a good idea
- Try cutout and mixup especially for small classification datasets







Step 1: Check initial loss

Turn off weight decay, sanity check loss at initialization e.g. log(C) for softmax with C classes



Step 1: Check initial loss
Step 2: Overfit a small sample

Try to train to 100% training accuracy on a small sample of training data (~5-10 minibatches); fiddle with architecture, learning rate, weight initialization

Loss not going down? LR too low, bad initialization Loss explodes to Inf or NaN? LR too high, bad initialization



Step 1: Check initial lossStep 2: Overfit a small sampleStep 3: Find LR that makes loss go down

Use the architecture from the previous step, use all training data, turn on small weight decay, find a learning rate that makes the loss drop significantly within ~100 iterations

Good learning rates to try: 1e-1, 1e-2, 1e-3, 1e-4



Step 1: Check initial loss
Step 2: Overfit a small sample
Step 3: Find LR that makes loss go down
Step 4: Coarse grid, train for ~1-5 epochs

Choose a few values of learning rate and weight decay around what worked from Step 3, train a few models for ~1-5 epochs.

Good weight decay to try: 1e-4, 1e-5, 0



Step 1: Check initial loss
Step 2: Overfit a small sample
Step 3: Find LR that makes loss go down
Step 4: Coarse grid, train for ~1-5 epochs
Step 5: Refine grid, train longer

Pick best models from Step 4, train them for longer (~10-20 epochs) without learning rate decay



Step 1: Check initial loss
Step 2: Overfit a small sample
Step 3: Find LR that makes loss go down
Step 4: Coarse grid, train for ~1-5 epochs
Step 5: Refine grid, train longer
Step 6: Look at loss curves



LOOK AT LEARNING CURVES!



Losses may be noisy, use a scatter plot and also plot moving average to see trends better









time



Loss



time



















Step 1: Check initial loss
Step 2: Overfit a small sample
Step 3: Find LR that makes loss go down
Step 4: Coarse grid, train for ~1-5 epochs
Step 5: Refine grid, train longer
Step 6: Look at loss curves
Step 7: GOTO step 5


CROSS-VALIDATION "COMMAND CENTER"

Instant States												
auther 1	worker 2	sector 1	souther 4	Servicer 3	ana kate di	Section 7	active 1	wighter 3	author 11	-	analas 12	workers 13
Tana anna 1911-191	Lord, and the second se	Anni aparamati, 2013-20- tempetativa na Anna a Manana S. SALANA Linea N. SALANA	Last symmetry 2011-50- Lynamic 20, 100000 Wine 10, 14, 100	Lori anticoli 2011-01- controlo di antico munici i 10100 Line 2002/0000	Contraction (Contraction)	And Andrew States	Larr months 201-01-	Ante antionnelle (1011-11)- composition en conception manifestation (1010-11)- temen (1010-11)- temen (1010-11)-	And printing Statistics		Last menters (0)(2 centration and examin- ments 2.542/00 Last 2.5574(2)000	Last metrical NED-D+ Letters on Dr. Prima metric Dr. Heller
Conception 5 International Conception 1, National Con- tions, Conception Conception (Con- Englishment Conception (Con- Eng	 Toron, real-real-real-real-real-real-real-real-	[1] Space (1) (2017) Anno participation (2017) https://www.space.org/ 1000/2017) Anno participation (2017) Anno partici	 Yesser, et al., e	 Solo, 11297, 115 Prov., polici, etcologitativa (j., Passantella Providentella, Sologitativa (j., Passantella Providentella, Sologitativa (j., Passantella, Providentella, Sologitativa, Sologitativa, Providentella, Sologitativa, Sologitativa, Providentella, Sologitativa, Sologitativa, Providentella, Sologitativa, Sologitativa, Providentella, Sologitati	C. Naud, 2010/P. D., Press, and a secondaria's in: Automative Control of Computational Activity (Control of Computational Activity) (Control of Computational Activity) (Control of Computational Activity) (Control of Computational Activity) (Control of Control of Control of Control (Control of Control of Control of Control (Control of Control of Control of Control of Control (Control of Control of Cont	 Non-Chapter S, Non-Chapter S, Non-Chapter S, Non-Chapter S, Non-Chapter S, Non-Chapter S, No. Non-Chapter S, N	[1] Yang Yang Yi, Wu Yong Yong Konstant's Lu, Yanasatt' Yang Yu, Yang Konstant's Lu, Yanasatt' Yang Yu, Yung Konstant's Lu, Yang Konstant's Lu, Yang Konstant's Lu, Yang Konstant's Lu,	 Aller (1997) 34, Aller (1997) 34,	 Apart (10)27 S. Print, ends, encoders?s G., Nassaan?s Print, ends, encoders?s (10)27 Print, Status, Status		¹ Number of Statement's Statement's Statement's Statement of Statement's Statement of Statement's Statement of Statem	 "Statution of the state of the
Nan Jacob (S) Secol (Secol (S)) (S), "Support S(200), "Secol (S) S(200), "Secol (S)	 New Jacob 19, 19, 19, 19, 19, 19, 19, 19, 19, 19,	Paul, South*: 55, Press, State, State: 105, Paugeon*: Plana*, Taintenart: Taint, Plana, St., State, State Paul, Press, St., St., St., St., St., Paul, Paul, Paul, St., St., St., St., St., St., St., St.	 Near Jacobins (N. S.) 	Party acciding access in San month's lo- Send access's lo- Send access's Co. "Support Party accession, "parameters" "sup", Party accession, "parameters", "sup- parameters, accession, so-	Para Jacobin III. Press, Incolo, and S. 201, "Support" Print Press, Television (Context, Print Press, Print	 Pass Journam In. Pendi Antoli Steri Lilli, "Saapoul's Taxad, Taxonizat's Hand", Taxad, Taxonizat's Hand", Pass Journal J. 	Nacional State (1997) (National States (Science) States (Science) Science (Science) S	Para, Jacobsen, The Paral, Jacobsen, Lander, Lin, "Tappent", Paral, Sciences, Phys. Rev. P. Paral, Phys. Rev. D 1, Paral, Paral, Phys. Rev. D 1, 1011011, Paral, Phys. Rev. D 1, 101101, Paral, Phys. Rev. D 1, 10110, Paral, Phys. Rev. D 1, 101		 Nasi Japanin'i Su, Seni Josephen'i HL, Passenin'i Tanin', Sananin'i HL, Passenin'i Su, Sanin', Sananin', Su, Sanin', Sananin', Su, Sanin', Sananin', L, 	Para (postar) (b) Presi, Intel Scott, 101, Vansen/S Plant, Spranster/S (Sol Alarty), Plant, Scottario (Sol Alarty), Plant, Scottario (Sol Alarty), Plant, etc., Sol (Sol Alarty), Sol
Netta mangota ga tizence" 11, 1920g tara" 1,000 Net Peri net" 1, 1000 (1000 201, 1000 (100") 1, 10, 1000 (100") 1, 10, 1000 (100") 1,	Particle inclusion and participation of the second	Nexts (Sectorist processor) 11, "diagonaly (Letter) Next Sectors(L), "Grades and" 21, "Next Sectors(L), "Grades and" 21, "Next Sectors(L), "Grades and "	Netter (nethiolog) and characteristic in, "state and" lock, here and most lock, history con-	Paul, enclosed 1, Second compared pro-constraints (1), Second condition (1), Paul (1), Second (1), Second (1), Second (1), Second (1), Second (1), Second (1), Second (1), Second (1), Second (1), Second (1), Second (1), Sec	Partie contained and connects and in the part of the first contained to the first contained to the second t	 Partie deviated (a) interested? Partie rest? 1215 Partie rest? 1215 Partie rest? 12 Partie rest? 12 	Partie characterized per construction in "disary near" in the tree level near" in "disary near" etc. Society of a	Particle development and dependent in "Minageneral" in the base inter-second in a finition, can't be an an an and the second second second second second second second sec	Party respect at the second the start start to the free and contain the second second seco		Nervo (derige) of gas (derived) ¹ (b) "sking (deriv" (b) b) (b) "sking (deriv" (b) b) (b) "sking (deriv" (b) b) (b) "sking (deriv (deriv) derived) ¹	Antice Association (p), threads (P), (A), Adding parally (A), Anno Antig parally (A), Machine (a), (A), Name (A),
Print De la contra contra company desertado a contra contra contra contra con	Provide Constraints and the second se	A weat in the point of the second sec	Press Conception on team and the control of team plane and the team plane and team plane and team plane and the team plane and team plane a	Peterser characteristics discretion/1- 1-area (1-area) (1-area) Recording (1-area) Peterseristics (1-area) Provided (1-area) Transition (1-area)	Press, Constanting on Copper- ment (Set Constant), Press, South Constant, Press, Press, South Const	President and a second	Production and a conservation of the second	Press, Chapter and Super- metric and Super- transmission and Super- transmission and Super- Transmission and Super- Transmission and Super- Superson (1), and Super- Superson (1), and Super- transmission (1), and Super- Super- transmission (1), and Super- Sup	¹ Cont., C. Stammeller and South Resolution (Darks) ¹ , ¹ State and South South ¹ State and South South ¹ States and South South ¹ States and South South South ¹ States and South South South ¹ States and South South South South South ¹ States and South S		Press Contractory and sugger manufactory and a second sec	Press, v. Scarnin er inger Bestilter steller, Statt som steller, S. M. Passing ster's 1, Millel, Passing of angen, stratter's
Construction and Construction Construction Interactions (Construction), Stationary Construction, Stationary Construction, Stationary Construction, Stationary Construction, Stationary,	 A stand or disparaty from states, description of C. Marganity and about the Neuroperatory interaction of the Neuroperatory and the Neuroperatory Neuroperatory (Neuroperatory) Neuroperatory Neuroperator	 Second and Report of Texas and Records and Constraints, Texas and Second Second Second Second Second Second Second Second Second Second Second Second Second Second Second Second Second	 Provide the deputy of the stream. Record and the stream of t	 Principal D., Stepst., Instantory*: Principal conferences in the second system instantoine (197), "In post metrics in Participation (197), "In post (197)," (1), Participation (197), "In post (197), "In the Participation (197), "In the second system", In the Participation (197), "In the second system (197), "In the second system (197)," In the Participation (197), "In the second system (197), "In the second system (197), "In the second system (197)," In the second system (197), "In the second system (197)," In the second system (197), "In the second system (197), "In the second system (197), "In the second system" (197), "In the second system (197), "	Constitution of the second	 Provide sectore state of the sectore sector sector sectore sector	 Construction of the second seco	Py construction of the second system of the seco	Construction of the second		 Construction of the second seco	 Statuti (Constraint) (Statuting)
Annu and an annual for the state	Among your anomalies in a 1 time per lanced 11-102e march front per lance 12-083	Paragraphic second Processing Same proc. Methods (1,4716) Same proc. Methods (1,4716)	Prove and American State (1996) Same per Methods (2006) Same (2006) Same (2006)	Property form perception (1997) Property for the p	Stang per Sector Comparison (Comparison (C	Theory and an address of the state	Server and Annual Print State	Annual and Annual and Annual and Annual Annu	Hanny Person annual Person (1997) Hanny Person		Compared provident of the second seco	Sengine C. Sengine (1977) 111
-	-	5	1:	- Unestadore	2	-	T + c	3	-		8	3
<u></u>	-		-	-					T	-	The second second	
analase 12 Internet (12)-to-	Acceleration 11 International and a second	Services 18 Innar Certe Last Mercenti III Indi- Letter Merca Jacob	unidate 17 ratio and last units Diff-to- rescention decar	Section 18 may lyin ber section 2019-09- lyin transmission	angelene 19 ministration tant partenel Ditt-Th- tantation(ta.outma)	worker 32 name and last spinster 305-11- control to control	anniher II. Sony losis San modele Dill-Di- Lencon in event	atolice 23 news acts and animal (RCh-1)- articles is animal	author 23 marchine 23 marchine 2010-12- marchine 10, march	weather 24 Visconstant 2010-11- Last weather 2010-11- Designation (Control	atalier 29 Constants on Les anistes 20-0-0- constants to be	anation 28 main 200 less and 2019-11- less constant and 2019-11-
port 1.14.00 For HERITEDIO Inc. 1414 (2004) Inc. 1414 (constants) (c), National Nation, Phase established (c), National Nation, National (c), National National (c), National (c), National (c), National National (c), National (c), Nati	Income 124-242-2000 Free 124424/202000 Free Provide	Second 29, 84/00 Files WHLIGHT 51 [Topol (1997) 51 Topol (1997) 51 To	Handt 21, 12/20 Vier 70011/06/20 1. ************************************	International Statements (Control of Statements) (Control of Statements (Control of Statements) (Control of Statements (Control of Statements) (Control of Statements)	New WORLDONG 1 Total (1997) 191 Prov. Color Statistics 1, 5, 5, 5, 5, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10	Income and a second sec	Hands LD, BUM Line WOOL 20200 1. Noted, 103(27110), 1. Noted, proceeding in the Spectrum Pro- Noted, "Tanak Antonial (1997), Noted, "Tanak Antonial (1997), Noted, Tanak Antonial (1997),	Second 22, MCM American WebColdCOM 1 April (1997) 12, Proceeding and an anti- processing and an anti- anti	hands & 1674 histor TWO (16170) 1. Spectra (1997) 10, Print, with presentative (1, "Statement") Print, "Statement (1, 1, 1997)	Result 5 2020 Max 802/5807 1 1904,01971 10; 110,01,010,0101011 1;; Patternet 111,010,02P;	Hands 2. Hold Mar. 2002/0000 1. Novel, prospersively, i.e., Parcenter's Prov., more proceeding's, i.e., Parcenter's Prov., more proceeding's, i.e., Parcenter's Prov., more proceeding's, i.e., Parcenter's Prov., more proceeding and	Brack M. 5210 Mar With 12100 1. "plant (1210) "rist, with president"s in Parcente" "rist", "large secondary star";
State and Antiparty 1.1.1. State and antiparty 1.1.1. State and antiparty 1.1.1. State Antiparty 1.1.1.1. State Antiparty 1.1.1.1. State Antiparty 1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.1.	Compared American Con- Name American Control of Con- Name American Control of Con- Name American Control of Control of Control of Control of Con- Name American Control of Control of Control of Control of Con- Name American Control of	 Anna area provided to the Anna area provided to the Anna area provided to the Anna Anna area to the Anna Anna Anna Anna Anna Anna Anna Anna Anna Anna Anna Anna Anna Anna Anna	 Song area, incomer¹/₂, L.S., Song area, incomer¹/₂, L.S., Song area, incomer¹/₂, S., Song	Paragener annuar's c.h. Paragener annuar's c.h. Paragener to the second	 Construction (1997) Construction (1997) 	 New Josephine Control (1995) New Josephine Con	Annual moders' 21. Annual moders' 21. Annual moders (1997) 100. Annual moders (1997) 100. Annual moders (1997) 100. Annual moders (1997) 100. Annual March (1997) 100.	 Vice, Vice, Jon County, V. J., Vice, and Records, V. J., Vice, and Strategy and Vice, Vice, Vice, Name 7, 66, Vice, Vice, 2007, 101, Vice, Vice, Vice, Vice, 2007, 101, Vice,	 Terris, and an observed in the result and secondary is an interval of the result and secondary in the result and secondary in the result and secondary is an interval of the result and result and secondary is an interval in the result and secondary in the result and result and result and result and result and	Press, Bar, Header V., etc., Press, Joint, Header V., etc., Press, Joint, Header V., Sch., Parriel, annihilitig, annul V., Hy- Rear, Barrielli, T., Bar, Parriel, Barrielli, S., Barry, B., Parrielli, B., Barry, B., Barry, B., Barry, B., Barry, B., Barry, B., Barr	Press, dos, manufer, star- ress, dos, manufer, sta- tures prof. annihility, annihility, Press, profile annihility, annihility, Press, profile 1, 15, Press, press, p	 Annu Participation (Control of Control of Contro of Control of Control of Control of Control of Control of C
Non pix in anit's to South, statist's 1, Notice interpret pix interestory 1, "Anity pixe" 1, 500, Non they pixe" 1, 500, Non they pixet	Part (10, 10, 40071) m; Name, restants (10, 10, 10, 10, 10, 10, 10, 10, 10, 10,	Non pt. 17 and 11 m. Franki, manual 1 L. Netter involution pt. Inconsect? Netter involution pt. Inconsect? Non Netter involution (Inconsect?) Non Netter involution (Inconsect?)	New years and the second secon	Perce pic, or places? - in; Practic relation? - i. Practic companies pic (transfer of); Practic company of pic (transfer of); Practic pic pic (transfer	March (M., M. 2007), Nr. Prantis, March (M. 1997), Nr. Marchan, Constanting (M. 1997), Nature (M. 1997), Nature (M. 1997	 Sant products and the second se	Para (dr. 17 deletti in), Paralle resoluti in Pretto cheratori dal transmetti 11, "Mara carto della della Pretto cheratori in della	Nami pin di anteri di Franci etternici in Nationa contratta dal disenteritti Internetternetti di di Francisco contra dalla	Para provid deserve inc. Press receptor to Press receptor to the Press receptor to the Press receptor to the Press receptor to the	 Arcelandor, Apparenter, Anne Anne pai, et glant/A, etc. Anne, Anne anne A, Anne Anne A, Anne A, Anne A, Anne 	Arministrative Agencies Control State and the description of Control Arministry of Control Control Control and Control of Control Control Control of Control of Control Control of Control of Control of Control Control of Control of Control of Control of Control Control of Control of Cont	Part pp. 14 AUDP 16 Parts, etc., and the first second seco
(1) Normal Annual Science and Annual Sciences (2010) (2	Fey, "Needer of "y 3: Needer of the second secon	http://www.org/st.tu Nacional.com/analysis.org/st. Complete complete complete Mercial contract of complete States (contract of complete complete States (contract of complete complete complete States (contract of complete complet	 Fill, "Seeking?", 1. Norman (Head (Head	[11] Yanakin W. Y. Yannar, Olivera Judge, Annuel P. Condition (Internet Internet Inter- Internet Internet), 2014, 47, 2014, 2014, 2014.	 Norman, Station of Stationary S	[15] Nonling Y. S. Namar, Johnson and American Pro- Press Contraction Complex Interaction Complex Pro- Interaction Complex Physics.	 Newton (PT, 1) Newton (Internet American)⁽⁴⁾ Newton (Internet American)⁽⁴⁾ Newton (Internet American)⁽⁴⁾ Newton (Internet American)⁽⁴⁾ Newton (Internet American)⁽⁴⁾ 	101, "Senior of 1.1 Normal Johnson (Senior Jonation)") Pressing of the senior of th	11.1. "Service of "1.5." Second plates, Judges, Accession," Second Computer Secondary Accession, Secondary, Computer Secondary, 1999, "Computer, Secondary, 1999," Computer, 1999.	Prot. (and, 2007), 1, "Voldan, control 51, "Parallelian", 5, "Voltar: José and State (Sciences)", "Control, Sciences, Sciences, automatic control, Sciences, automatic control,"	 Yoo, And Joseff, L. "Solder, Joseff, You, "Realistant", 5; Norise: probably solders, Accession", "Social Control of the Internation Internation (Internation) 	D.C., "Sendering", S. Harrison, Collection, and party advances of the Proceedings of the sender of the sender measurement of the sender of the Proceedings of the Proceedings of the sender of the Proceedings of the Proceedings of the sender of the Proceedings of
Constant of the second se	Construction of the second sec	Constraint, estate, discrimination international constraints and and a second constraints of the second constraints of the second constraints of the second constraints of the second constraints of the second constraints of the second constraints of the second constraints of the	 Constraint is used at an entry (- Constraint is used) and a second of the maximum of the second of the second of the second second of the second of the second of the second second of the second of the second of the second second of the second of the second of the second of the second second of the second of the second of the second of the second second of the second of the second of the second of the second second of the second of the second of the second of the second second of the se	 Constant II, and an international sector of the sector of t	Commercial and an entropy of the second seco	 Annual Annual March Meeting ** Constant of Annual Meeting ** Constant of Annual Meeting ** Annual Meeting ** Annual Meeting ** Annual Meeting ** Annual Meeting ** 	Comparis, maps, strettery: community of the part distribution manual distribution (1), and being and distribution (1), and being and distribution (1), and being and and an and an and an and and being and an an an and an and an and an and being and an an an and an an and an and an and being and an	 Construction of the second seco	A second	Comparis, angle, monthly ' comparise angle, monthly ' comparise of the second	 Construction, American Structure, Construction and American Structure, Construction and American Structure, International Conference on Construction International Conference on Constructio	 Construction of the second seco
Samola and Lark, Scient, Samola, Samo particult 1.0, Samola production of the samola statement of the law provide and statement samola constants and statement	America and a set of the function comparison for the set of the set of the comparison of the set of the set of the comparison of the set of the set of the comparison of the set of the set of the set of the set of the set of the set	 America and Science, And Science, Support, America property (1975). America processing Press, 1976. America and America Press, 1976. America and America Press, 1976. America Press, 1976. America Press, 1977. America Press,	 America and Const. Proceedings. Computer, Name particular Con- transportation and Proceedings. Computer Systems 2014;2014 Computer State State	Personal apply and, "entropy", "employed, "end person", total "employed, account?", total law personal account?", total law personal account?", total personal total and total	Control, and C. Serli, "Antipart", "Support," Society and Serling, "Antiparticle and service of the production of the service and the service of the service and the service of the service of the service and the service of the service of the service service of the service of t	 Summary, and Savery, Save	Terrori, early lard, Sacherfy "Sequence", Security and the Heat production of the large production of the large transformer	 Among and Lord, Nation's Second, New particity (1), Second products of the land transmission of the model transmission of the 	Perspectral, News, personal v. Lin, https://www.asconfiller.com/ home-per-sector advectory personal vector and v. G. MT NE ATL	 Constant, and S. Serli, "Anti-endi- transprop", Neuro, pathod?: In Phys. Rev. B 100, particular 71, 514 (1999). Internet Constant Phys. Rev. B 100, particular 2016. 	 Territo qui'i trell, festerafic tonanafi, fanai partiafic trali, renn princamatific trali tone per lambdull, dile menter (state qui etc.) 	Personal application, forceworks responsible, force particular, force response processing of the particular particular processing of the particular particular processing of the particular particular particular particular
		- +		-		-		-	ELEL	11.000	14.345	-
-				i i					E. The	-		
striker 13 inter see	worker 18	worker 28 note one.	worker 30 team size	seriar H new spa	worker 32 new une	Rocker II response on the universe description	atolar 14 and and anter article	stering 35 monto constante averante	worker 34	etchar 37 noter other care systems beindle	worker 38	worker 39 new men
unter disease and a second prior of a second s	Control of the PACTURE Space for PACTURE Prior Paceto Action of the Paceto The Paceto Action of the Paceto Acti	Contraction of Annual States of Annual Annual States of Annual Annu	Contraction of Amount and the Advance and Contractaneous (Amount and Amou	And A Contract of the second s	Production of the second secon	PTERSON IN DECEM- MARK STREET INFORMATION INFORMATION TOTAL AND ADDREET STREET	And the second s	CTUDEN ALTON MAR TRUCKSON LAN MULTINE LAN ALTON LAN	L. COLUMN AND ADDRESS MARK IN ADDRESS	A CONTRACTOR OF A CONTRACT A CONTRACTOR OF A	ACTIVATION DI LI ANTO Martina di Martina Inter Antonia di Antonia Inter Antonia di Antonia di Internazione di Antonia Inter Antonia di Antonia di Internazione di Internazione di Antonia Internazione di Antonia di Internazione di Internazione di Antonia Internazione di Antonia di Internazione di Internazione di Antonia Internazione di Antonia di Internazione	And Antonio antonio antonio antonio antonio antonio antonio antonio antonio an
 Yend And Mappins, -L. Sang print Animal You Sub- band annalising starts, Map. New Annual Sciences, 1995. Sang Animal Sciences, Nature Sci	111. "real and tempt"L. "step proc. Bullet": 1.4 "step proc. Bullet": 1.4 "step proc. Bullet": 1.6 "step sector", 10. "step sector", 10.	 Yessi and Septer's 'L' New Sectors Sectors's Sol, New Sectors Sectors's Sol, New Sectors's Sol,	 Yerni, Ann. Statest', -L. Yerni, Ann. Statest', S.A. Yerni, Statest', S.A. Yerni, Annual's Statesting, Statest', 1985, Yerni, March Y., 198, Yerni, March Y., 198, 	 Weig and State State C. 11. Weig and Schladt St. 11. Weig and Schladt Schladt St. 11. Weig and Schladt St. 11. Weig and Schladt St. 11. Weig and Schladt Schladt St. 11. Weig and Schladt Schladt Schladt St. 11. Weig and Schladt Sch	 Series, Series, Stanford, M., Sarah, K., Marchard, S. & A., Pranti, Sciences, Street, S. Bill, Pranti, Sciences, C. & Stanford, Nature, Sciences, Conf., Sciences, Sciences, Nature, Sciences, Conf., Sciences, Sciences, Nature, Sciences, Conf., Sciences, Sciences, Nature, Sciences, Sciences, Conf., Sciences, Nature, Sciences, Sciences, Conf., Sciences, Nature, Sciences, Sciences, Conf., Sciences, Nature, Sciences, Sciences, Sciences, Sciences, Nature, Sciences, Sciences, Sciences, Sciences, Nature, Sciences, Sciences, Sciences, Sciences, Sciences, Nature, Sciences, Sci	Transp. extending, state?, 100, Transp. Jone, State??, -1, Transp. prod. annume??, 11, Transp. prod. annume??, 11, Transp. prod. annume??, 10, Transp. prod. 7, 10, Transp	 Testi and testing of the best prior best testing best sectors and the best sectors and the best sectors of the be	 Vent and import of a range produced market in a range produced market in range ma	 Terris, and Instant's -1, then proc. Ballant's and Provide scotting scott's Bills, New Sector's Dig Provide scotter's Dig Proc. Proc. Ballant's Dig Proc. Ballant	 Constanting Constants, 1997,	CL, New Ant Depart - L, The plan million - 1 - 1, New position methods - 1 - 10, Part model - 10, Part Math. 10,	File, "real and result": -1, "sing point backed "r a d "serie mention score": into the weather into the weather into the weather into.
George, Appleasantes Franks, Non-patrice above 1, 20, Madie plantantes, patrice and the Matter memories and patrice and the OL Microsophics (20, 2017)	Ricel, Semester's Casel, Pers, pr., or, description, Packer, Station, C., Packer, Station, pp., semester's 21, "Security Station,	Parton, Parmineral's Name, New Job, or destrict the Parton Research to New York, memory and stressolution (1, "deservation") 1,00%	Partoli, Spitzenszi S. Sandi, Nan pol. or aborti S. St. Paulo Science S. Partol Science and Americanic Discourse and Science Sciences 11, Spitzer and Sciences	FaceF, Satisser's SateF, New Jac A. Alexi's 10, "SateF, SateF, 11 Partie (SateF, 1) Preve (SateF, 1) (1, "Biological") 1,6,	Parie and research frame. Parie and research for the Parie constraint for Parie constraints and descending \$1, "Sing staff, U.S.S.	President and Still, Nagandy Statistical Constants (Statistical Statistical Constant), Statistical Statistical Statistics (Statistical Statistical Statistical Statistical Statistical Statistical Statistical Statistical Statistical Statistical Statistical Statistical Statistical Statistics	Caref, Spinnard's Funds, New Jack of America St. Nation, Company, and Science St. Nation, company, and strength St. Di, Nation, combined by Jack	Fixed', Springers's Fords', See and in America ID, Spring Springers, 5, Spring American and American Discourse and American Discourse and States,	Flands, Salananashi Kalads, Nan ppi, ni akushi Ib, Maslir, Maslani Ib, Pretin, Handari I, aku mentelahi Ib, Mesajadahi 1,00,	Cont. Partnerstar's Galary Para part in description, Paralle classification, part Paralle classification, part internation (1, 1960), parallel (1997).	Faren, Apparature Cataly, New Joint, et al. (2017), 35, Partie (Artificial), 5 Parties (Artificial), 50 Parties (Artif	 Faref, Santaspere's Garafy Santaspere as Annihi Sh. Santashi Santaspere as Santashi Santaspere as demaintaile Santaspere as demaintaile Santaspere as demaintaile
(1), "Annual of a line of a second state of a	All, "research of 1.2. Persons states conservations": "conservations": Research of the servation of the s	All, "Researce"s 5. Particle Andrew School, Ministruc"s Constitution and an and a second second	 Statement of York, Construction Pro- Science Learning and Sciences Pro- Provide Learning and American Pro- Resources International Construction Pro- Resources International Constru	 Statistical Statistical Control of Control	All, "respirat", 1, None and respiration and the result of the second se	 Construction of the second seco	 State and State and Sta	 Statement of V. L. Statement of V. Statement of V. S	 Annessento C. S. Schemerson A. S. Schemerson and Schemerson A. S. Schemerson and Sc	 Statistical Statistics Statistics Statis Statistic	 Network (Constraint) 	 Construction of the construction of the construction
Saarding (2007) (colors) Social and a second second sec- social conference in the second second conference in the second	 Standburg, Schröfel, Landong, Schweissellung, andreast, Microsoftwarfel, - Schweissellung, Schweissellung, Schweissellung, Standburg, Standburg, Martiness, Schweissellung, Sc	Association (Association) Association (Association) Construction (Association) Association (Association) Association (Association) Press, construction (Association)	Association, closely, distance, for Astronomical content, discontract, for association of the state of the manufacture of the state of the State of the state of the state of the state of the State of the state of the state of the state of the state of the State of the state of the state of the state of the state of the State of the state of the state of the state of the state of the State of the state of	Presenting, 2010, 1 - 2010, 1 Advantage of Annual Annual Annual Annual Present of Annual Annual Annual Annual Annual Annual Annual Annual Annual Annual Present Annual Ann	Providing (2007) substants Provident A. Antonio Scientification Provident A. Antonio Science (2007) Provident A. Antonio Science (2007) Pr	Principal prior, internet, Principal and a second second second metal second se	Presidence, March, S. Statelly, Proceedings, M. Schutz, Microsoft, S. Statelly, Construction, Comparison, Comparison, Comparison, Construction, Comparison, Comparison, Comparison, Comparison, Neurophysics, Comparison, Comp	Providing (2007), is attempt formulation or solution (attemption) in an interpretation (attemption) interpretation (attemption) Providing (attemption) (attemption) Providing (attemption) (attemption) Providing (attemption) (attemption) Providing (attemption) (attemption) (attemption) Providing (attemption) (attemption) (attemption) (attemption) (attemption) (attemption) (attemption) (attemption) (attemption) (attemption) (attempti	Paralling (2007) - Letting Processing of the state of the state of the state of the state of the state interview of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the state of the st	Presenting (2007) 1, million, Comparison, company, deserving (2) Construction particular company Construction particular company Construction (2), resp. particular (2), re- Present, construction (2), re-	Paralling (Alar), is served, Principality, angus, discovery (1, "received a deputy for impe- manual received and the data of the Paralling of the served of the	Property of the Content of the Conte
New Arrange Control of the second sec	Teacol, as and the "Stage", 1 (1999). Assessed paper (1999). Stage "Stage", Compared", "Strate, period" (1999). Stage parts assessed Pri 1 (1999). Stage parts assessed Pri 1 (1999).	Postoli accello (11, 10007) 1.000- b. Nations, party and, Nationaly Postorer, Next, presserve 1.02 National part (200007) 1.02 National part (2000077) 1.01 National part (2000077) 1.01	Count and a market and the provided and the count approximate and the count of the provided and the count of the the provided and the count of the the provided and the count of the the count of count of the the the count of the the the the the the the the	Presiduated's 10%, "beat"s a me- ta comment, agent analy, "accessive comment, "read, areasoft"s tota "reagents", "read, areasoft"s tota "reagents", and analysis of the read of the second of the	Parcilla scarf (10, "lags", 1, then Planting approximately formed (Propagately, freed, proximately) Planting apple, and proximately (from plan schement), resp.	Pascill et al's 105, "sould" y the- its, "allocit, part," sould, "annual", "sould all and annual", 121, "sould plate annual"," 121, "sould plate annual", 121, "sould be annual to be annual to be annual to be annual to be annual to be annual to be annual to be annual to be annual to be annual to be annual to be annual to be annual	Parall, and a pre-local, foreits for the based of the second seco	Theory and the Theory of Strength 1 Theory and the Strength 1 Theory of Strengt 1 The	Tracial scatt ² (10), Trace's Link- 1, "anistic query land, "anistrati- tic scatter,", "track and "state", Link- many plan anisolative (Link- tics), plan	Parcell, eccard 1, 69, "separt", 1, 169-1 Assessing agents since, Assessard 1, Parcell, and Assessing 1, 1, 19, 1 Parcell, prod. assessing 1, 1, 1, 1 Parcell, prod. assessing 1, 1, 1, 1 Parcell, prod. assessing 1, 1, 19, 1 Parcell, prod. assessing 1, 1, 1, 1 Parcell, prod. assessing 1, 1 Parcell, p	National Action (1999), National Constraints, National Constraints, National Constraints, National Constraints, National Science (1999), National Science (1999), National Constraints, National Co	Parint Jones, 1997, 2007, 2007, 1100- to, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2009, 2007, 200
	13.134	13.452	11.314	53.861	11.552	54.901		11 TH	59.049	11.945	12.343	51.402
	1	1	1	E		- Marine	1		11		-	E
nature 41	anches 41.	marker 43	walkes 43	antine 44	michae 45	marker 48	states 17	atabas 44	and an and a start	Riphur M.	Riches SI Transfer State	Richas II.
ente la constante note la constante de la constante la constante de la constante la constante de la constante de la constante de la constante de la constante de la constante de la constante de la constante de la constante de la constante de la constante de la constante de la constante de la constante del constante de la constante de	MONTON, JOSE work 12, 4700 March 001000700 1 Sectors 1 Sectors 1, Sectors	Province and concepts work # 32.000 Fair Mark/10008 1 North Concepts (*) 1, Tananak *	Weatherst.comp weath 12, 80/00 Fault VERSIANDER 1- Name (12, 97) 125, Name (12, 19) 125, 13, 13, 13, 13, 13, 13, 13, 13, 13, 13	Process in April 19 mode F. H.CH Final profiling Con- Profiling Conference on April 19 Profiling Confere	Inflamma ta, manyo menti 15, 14/18 Inflamma Inflamma Pana, bella recoderati 1, futuranti	Antonio II. GAN March 17. GAN Kasa Wan Karta I. Anna Jan Jan I. Tana Antonio Kasaka I.	Printer In. Print Sector 1. LUM Francescone Printer States I. L. Spears	Artisten H. Johnson Needs D. Note Hart Transform - State (1997) 21, "Artisten"	Provide Statements - Science meters 15: 34/14 Francisco - Science - Scienc	Protection La constant week 7,38/30 Lass control (constant) - State (constant) Peter control encoders*1 (L. *)procession	Antonio da Parte Media M. Mali Liai Antonio Mali 1 Antonio Mandersofi I., "Matane"i	Arthorney, Frank March 15, McHill March 16, McHill March 16, McHill Print million standard 1, 5, 75, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10
And Annual Annual Comparis of the second sec	 State of the second seco	 Construction of the second seco	 Prover, "Plaque attaching state", Plaque participation of the state of	 Print, "Print, part (assumption of print), Print, "print, part (assumption of print), Print, print, print, print, print, Print, print, print, Print, print, print, Print, print, print, Print, print, Print, print, Print, print, Print, print, Print, print, Print, print, Print, print, Print, print, Print, print, P	 Construction of the second seco	 State and State and Sta	 Paras, "One account, etc." Paras, and parasite "Con- trans, and account", 202. Paras, account, "Con- trans, account, account," 202. Paras, account, 202.	 Pringer, Annual Stream, 2010, 197 Pringer, 2010, 197 Pr	 Process, "Comp. Second and a state", Sci. Process, and the comparison of the state of the state	Financial Control of C	Providence control too, Providence control too, Providence control too, Providence control too, Providence control too,	Provide The Providence and the United States of the Providence and the Providence of
And Andrewski, and Andrew Andrewski, and Andrewski,	Page 1, application of the second sec	Parage, "parameters to the composition Parage and composition to the Parage and composition of the Parameters and the composition of the composition of the composition of the Parameters and the composition of the composition of the composition of the Parameters and the composition of the composition of the composition of the Parameters and the composition of the composition of the composition of the Parameters and the composition of the Parameters and the composition of the compositi	Conf., "planate", "conf. No. pt., it death, it, "conf. water", 1. Conf. water", 1. Conf	Aurora, Application Aurora, Au	Press, Spinster, State, State, Science, State, State, Science, St. State, State, St. State, State, St. State, State, St. St. St. State, State, St. St.	 Sandy, Spherosciet, Spherosciet	Parage, spatiantics, comments barry processing and second	Paral, "planter" "start" Terrar, "planter" "start" Terrar, "start", fa Table, "start", fa Table, "start", fa Table, "start", fa	Particle Systems of States	Paral Salar, Salar, Salar, Hagaar, Salar, Tamasan'i Salar, Tanagi si Ameriji Bi, Tana Ameriji Bi, Tana Salari, S	Parall places, and y ces, reasonable Faced, Appendicate's Cond's Note pair or dent? Ma Participations to Participations and reasonable	Paral Sares Alarty 10, "Sagard" Saret, "assessed", Saret, Nata pp., 10, 4000 (1, 10, "Saret, Saret, 1, 10, "Saret, Saret, 1, 10, "Saret, Saret, 10, 10, 10, 10, 10, 10, 10, 10, 10, 10
Yoon, Deed Social's C. "Robbins, close" and, "Hospital and C. Socialization and control of sectors of a research of sectors and sectors of the research of the sectors of the sec- tors of the sectors of the sectors." Inter- tional Conference of the sectors of	¹ Min. Rev. (1997) 1, 1 (Science, 1997) and "science, science, electrony", 5 Constant, Problem, Science, electrony", Constant, Constanting Science, Stager.	The last work is " to be one" and " second of the second of the second Contract of the second of the second of the second of the	The feet work's L. "Ander your" http://www.setter.com/ Setters.com/setters.com/ Commentation.com/ Commentation.com/ Commentation.com/	The Art Long 1, "Sume you") re, "summary 1," Prove strateging to summary " Prove strateging to sum-	The loss how'r L. Traine, che'r ne, "Anderse" C Panlae offic olan donerer'r Traine offic olan donerer'r	 Tax See, See", L. Tosse, sor", str. "scattering", J. Security Carlos, Science, Million (1975) "Security Carlos Content, Million (1975), "Security Carlos Content, Million (1975), "	Terris Contraction (C. Manana, according and, "According of the Provider John on Science, According (C. Provider John on Scien	Teles, Band Society, L., Norma, Carry, Sec., Names and C. Sectore, Johnson, Strands, Konstanty's Transition (Strangering) (Strangering)	Trans, Bard, South, S., Nalasa, Josef, and "Stationary S. Parcine Joints, Johnson, Schulter," Constant, Schulter, Schulter, Schulter, S.	 Marta and T. Like, Nuclear Josef A. Science (1997) M. Personnelli, International Problem Annual Advances/P. 	 Merce and Y. 1981. Marchael (2007) 1, "Marcal, 2007" 1, Marchael (2007) 1, "Marcal, 2007" 1, Marcal (2007) 1, "Marcal (2007). 	 Marcar, and Y. L. Hu,





RANDOM SEARCH VS. GRID SEARCH



Random Search for Hyper-Parameter Optimization, Bergstra and Bengio, 2012

GT 8803 // FALL 2018



TRACK RATIO OF WEIGHT UPDATES/ WEIGHT MAGNITUDES

```
# assume parameter vector W and its gradient vector dW
param_scale = np.linalg.norm(W.ravel())
update = -learning_rate*dW # simple SGD update
update_scale = np.linalg.norm(update.ravel())
W += update # the actual update
print update_scale / param_scale # want ~1e-3
```

ratio between the updates and values: ~ 0.0002 / 0.02 = 0.01 (about okay) want this to be somewhere around 0.001 or so



HYPERPARAMETERS TO PLAY WITH

- network architecture
- learning rate, its decay schedule, update type
- regularization (e.g., dropout strength)

neural networks practitioner music = loss function





SUMMARY

- Improve your training error:
 - Optimizers
 - Learning rate schedules
- Improve your test error
 - Regularization
 - Choosing hyperparameters

