## Lecture 11: Training Neural Networks (Part 2)

Justin Johnson

Lecture 11 - 1

### Reminder: A3

- Due Monday, October 14
- Remember to <u>run the validation script</u>!

### Reminder: Midterm

- Monday, October 21 (two weeks from today!)
- Location: Chrysler 220 (NOT HERE!)
- Format:
  - True / False, Multiple choice, short answer
  - Emphasize concepts you don't need to memorize AlexNet!
  - Closed-book
  - You can bring 1 page "cheat sheet" of handwritten notes (standard 8.5" x 11" paper)
- Alternate exam times: Fill out this form: <a href="https://forms.gle/uiMpHdg9752p27bd7">https://forms.gle/uiMpHdg9752p27bd7</a>
  - Conflict with EECS 551
  - SSD accommodations
  - Conference travel for Michigan

### Reminder: No class on Monday 10/14! (Fall Study Break)

Justin Johnson

Lecture 11 - 4



### 1. One time setup

Activation functions, data preprocessing, weight initialization, regularization

### 2. Training dynamics

Learning rate schedules;

hyperparameter optimization

### **3.After training**

Model ensembles, transfer learning,

large-batch training

**Last Time** 

Today

Lecture 10 - 5

### Last Time: Activation Functions



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



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



#### October 9, 2019

Justin Johnson

Lecture 11 - 6

### Last Time: Data Preprocessing



#### Lecture 11 - 7

### Last Time: Weight Initialization





Glorot and Bengio, "Understanding the difficulty of training deep feedforward neural networks", AISTAT 2010

#### Justin J<u>ohnson</u>

#### Lecture 11 - 8

### Last Time: Data Augmentation



Justin Johnson

Lecture 11 - 9

### Last Time: Regularization Cutout Training: Add randomness **Testing:** Marginalize out randomness **Examples**: **Batch Normalization Data Augmentation** Dropout DropConnect Fractional pooling Mixup X

#### Justin Johnson

#### Lecture 11 - 10

#### October 9, 2019

Stochastic Depth



### **1.One time setup**

Activation functions, data preprocessing, weight initialization, regularization

### **2. Training dynamics**

Learning rate schedules;

hyperparameter optimization

### 3. After training

Model ensembles, transfer learning,

large-batch training

Today

Lecture 10 - 11

## Learning Rate Schedules

Justin Johnson

Lecture 11 - 12

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



#### Justin Johnson

#### Lecture 11 - 13

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



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

#### Justin Johnson

#### Lecture 11 - 14

# 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

#### Justin Johnson

#### Lecture 11 - 15

### Learning Rate Decay: Step



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



#### October 9, 2019

#### Justin Johnson

Lecture 11 - 16

### Learning Rate Decay: Cosine



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", ICCV 2019 Radosavovic et al, "On Network Design Spaces for Visual Recognition", ICCV 2019 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: 
$$\alpha_t = \frac{1}{2} \alpha_0 \left( 1 + \cos(t\pi/T) \right)$$



#### Justin Johnson

Lecture 11 - 17

### Learning Rate Decay: Linear



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

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

Devlin et al, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", NAACL 2018 Liu et al, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", 2019 Yang et al, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", NeurIPS 2019

#### Justin Johnson

Lecture 11 - 18

### Learning Rate Decay: Inverse Sqrt



**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: 
$$lpha_t=rac{1}{2}lpha_0\left(1+\cos(t\pi/T)
ight)$$
  
Linear:  $lpha_t=lpha_0(1-t/T)$   
Inverse sqrt:  $lpha_t=lpha_0/\sqrt{t}$ 

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

Justin Johnson

#### Lecture 11 - 19

### Learning Rate Decay: Constant!



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

Inverse sqrt:  $\alpha_t = \alpha_t$ 

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

Constant:  $\alpha_t = \alpha_0$ 

Brock et al, "Large Scale GAN Training for High Fidelity Natural Image Synthesis", ICLR 2019 Donahue and Simonyan, "Large Scale Adversarial Representation Learning", NeurIPS 2019

#### Justin Johnson

#### Lecture 11 - 20

### How long to train? Early Stopping



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. **Always a good idea to do this!** 

#### Justin Johnson

#### Lecture 11 - 21

Justin Johnson

Lecture 11 - 22

### Choosing Hyperparameters: Grid Search

Choose several values for each hyperparameter (Often space choices log-linearly)

#### Example:

Weight decay: [1x10<sup>-4</sup>, 1x10<sup>-3</sup>, 1x10<sup>-2</sup>, 1x10<sup>-1</sup>] Learning rate: [1x10<sup>-4</sup>, 1x10<sup>-3</sup>, 1x10<sup>-2</sup>, 1x10<sup>-1</sup>]

Evaluate all possible choices on this hyperparameter grid

### Choosing Hyperparameters: Random Search

Choose several values for each hyperparameter (Often space choices log-linearly)

#### Example:

Weight decay: log-uniform on [1x10<sup>-4</sup>, 1x10<sup>-1</sup>] Learning rate: log-uniform on [1x10<sup>-4</sup>, 1x10<sup>-1</sup>]

Run many different trials

### Hyperparameters: Random vs Grid Search



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

Justin Johnson

Lecture 11 - 25

### Choosing Hyperparameters: Random Search



Radosavovic et al, "On Network Design Spaces for Visual Recognition", ICCV 2019

Justin Johnson

#### October 9, 2019

Lecture 11 - 26

(without tons of GPUs)

Justin Johnson

Lecture 11 - 27

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. Turn off regularization.

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 learning curves

### Look at Learning Curves!



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

#### Justin Johnson

#### Lecture 11 - 34



Just	in J	0	hn	son

Lecture 11 - 35



		• •							
	L C	ГΙ	n		าท	n	C	$\cap$	n
J	US.	LI		JU	/		2	$\mathbf{U}$	

Lecture 11 - 36


Lecture 11 - 37



Lecture 11 - 38



Lecture 11 - 39



#### time

JUST	tin J		SOL

Lecture 11 - 40

## Choosing Hyperparameters

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

## Hyperparameters to play with:

- network architecture
- learning rate, its decay schedule, update type
- regularization (L2/Dropout strength)

neural networks practitioner music = loss function



#### Justin Johnson

Lecture 11 - 42

### Cross-validation "command center"

and a second sec												
antibut 1 Info general, BLI-Da- ment 18. NUM Ten 1900/00/190	workne 2 net inte ter antenis 2011-00- terstructure on the work 20-15-00 ter 10-15-00-00-	Section 1 int sectors int sectors int sectors int 54000 int 54000	Westlerin 4 Testi antis- lari aprinisti 2011-20- trati da di constituti Westleri 2014/201000	Security 3 Loss and 101-02- loss and 101-02- loss of 02 and 02 New 101000	Antoine 6 Terrison and Terrison and the second se	Section 7 intermediate 2015-01- intermediate 2015-01- ment 20.7009 intermediate 20150	Anteriory 1 Sectors (1997) Sectors (	Mighter 3 mighter 3 international (021-11- international) (021-11- international) (021-11- international) international (0210)	marries 12 marries 12 http://www.science. marries.com marries.com htm://www.science. htm://www.science.		analyse 13 tan units into the pro- perturbation into the pro- perturbation into the pro- late Architeme	Augebra 13 magniness and less and and 122-11- less and an even magnine in Media less product results
L NOR, HUPL 5, HULL, NE, STRABERT 5, S. Nataratti Solid, "State and and S. S. Nataratti Solid, "State and Solid State Trans." Solid State and Solid State Trans. Solid State Solid State State Solid Solid State Solid State Solid State Solid State Solid State Solid State Solid State Solid State Solid State Solid State Solid State Solid State State State Solid State State State State State State State State State State State State State State State State	L. SPORTON, U.S. S. Samuel, N. Samuel, S. Samuel, S. Samuel, S. Samuel, and S. Samuel, and S. Samuel, S. Samuel, and Samuel, a	[1] Spins, (1975) 1; Anno. edit provident's 3: A Statustic's Society, "Search provident scient", and, "Society Constraint", ed. Society, and public Society, ed. Nucl. Society, and Society, ed., Nucl. Society, and Society, ed., Nucl. Society, ed., 51, Nucl. Society, ed., 51, Nucl. Society, 65, Nucl. Society, 65, N	<ol> <li>Spectra (1997) 115</li> <li>Neto (1997) 116</li> </ol>	<ol> <li>None, 1939, 1939, 10, Prome price memberships, Passanethy Prome price memberships, Passanethy Prome, and an annual statements of the Press, and Annual Statements (Statements), 1935, Press, price memberships, 2019, 1935, Press, 2019, 2019, 2019, 2019, 2019, Press, 2019, 2019, 2019, 2019, Press, 2019, 2019, 2019, 2019, Press, 2019, 2019, 2019, 2019, Press, 20</li></ol>	<ol> <li>Non, 10 (P. O.).</li> <li>Non, 20 (1977) Oc.</li> <li>Non, 20 (1977) Oc.&lt;</li></ol>	<ol> <li>Neuropean St. St. St. Stranger, N. S. Stranger, S. S. Stranger, and S. Stranger, and S. S. Stranger, and S. Stranger</li></ol>	C. News, 1997; Phy. Y. H. Strang, Stranger, M. S., Samana, N. Yani, "News, News, News, 1997; News, 2008, 2008; News, 1997; News, 2008, 2008; News, 1997; News, 2008; News, 2008; News,	<ol> <li>Company and Annual State (1997)</li> <li>Company</li></ol>	<ol> <li>Name, et al. (* 5).</li> <li>Parameter and a procedure (* 6).</li> <li>Parameter and (* 6).</li> <li>Parameter (* 6).</li> </ol>		<ol> <li>Nyani, Arabini Su, Yang Ma, Kanang Minaka, akitifi Narih, Yang Minaka, akitifi Marabana, akitifi Sang Ang Ang Ang Ang Ang Sang Ang Ang Ang Sang Ang Ang Ang Sang Ang Ang Ang Ang Sang Ang Ang Ang Ang Sang Ang Ang Ang Ang Sang Ang Ang Ang Ang Ang Ang Sang Ang Ang Ang Ang Ang Ang Sang Ang Ang Ang Ang Ang Ang Ang Sang Ang Ang Ang Ang Ang Ang Ang Ang Sang Ang Ang Ang Ang Ang Ang Ang Sang Ang Ang Ang Ang Ang Ang Ang Ang Ang Sang Ang Ang Ang Ang Ang Ang Ang Ang Ang A</li></ol>	<ol> <li>Charles (1997) Mr.</li> <li>Yons, and a monotone's 'Le "assessmith's final", "Final Activity (1998) Mr. (1997)</li> <li>Weit, "South, Monotone's Le (1), "Name, South, Monotone's Le (1), "Name, South, Monotone's Le (1), "Name, South, Monotone's Le (1), "Name, South, Market South's (1996)</li> <li>Name, South, Mr. (1996)</li> </ol>
Secola Sciences (1995), Magnetolia Colord, Contention (1996), New York, New York, New Parallel, Sciences (1997), New York,	Presidential start (1996), Transmith Control, New York, Annual (1996), New York, Annual (1996), New York, Annual (1996), New York, Annual (1996), New York, New York, New York, New York, New York, New York,	Densi, Jassis, Jonef's, Olis, Pangenet's Classify, Agencientet's October, Name, Proceedings, 1999 Panelli, Proceedings, 1999 Panelli, Proceedings, 1999 Panelli, Proceedings, 1999 Panelli, Panelli, 1999 Panelli, Panelli, 1999 Panelli, Panelli, 1999 Panelli, Panelli, 1999 Panelli, Panelli, 1999 Panelli, Panelli, 1999 Panelli, 1999 Panelli, Panelli, 1999 Panelli, 1997 Panelli,	New Joseph Joseph 100, New Joseph Treed, New Joseph 10, New Joseph Namil, Press, N. S. Strand, Namil, Press, N. S. Strand, Namil, Press, New Joseph 10, New Joseph 10, New Joseph 10, New Joseph 10, New	has seen to be been been been been been been been	Detail Josep, prod 1, 202, Nappent's Closel, Nappentier's Mo. Closel, et al. 2017 Mo. Closel, et al. 2017 Mo. Closel, et al. 2017 Mo. Closel, and an and second St. 21, March 2, 2017 Mo. 2018	Densi, Janeis Scand's UK, Vangenet's Care, Stransmither & Hand's New Joy, A. Janeitter M. Mark, Constanting and Constanting Careful Annual Constanting Con- Net On Constanting Conf. (Constanting) 11, 1981, 2007;111(2).	Period States, Scott St., Passent's Pariot, Taking Scott St., Pariot, P., Scott St., Pariot, St., Scott St., Pariot, St., Scott St., Pariot, St., Scott St., Pariot, St., St., Scott St., Pariot, St., St., Scott St., Pariot, Scott St., Pariot, St., Scott St., Pariot, Scott St., Par	Presidence, statistics (St., Stagenet's Control, Statistics), Statistic, Statistics, et al. and St. Statistics Statistic constant and statistics (St. St., Statistics), and Statistics (St. St., Statistics), and Statistics	Densil James Analys (M., Yangarati's Const., Yanasanari's Mark, Yana, Jawatani K., Katali, Yana, Analysis, J. (2019) Participation of the Constant Processing States of Social Sciences of Social Sciences (No. 1997) 11, 1997 (2019) (2019).		Period Sector, econity, 2016, "Company"s Transf, "Sectorerstore"s "State", "And gas, etc., according to Parally, etc., according to Parally, etc., according to Parally, etc., according to Parally, a	<ul> <li>Presi, Janas, starbi 201, "Sagaraphi Plant", "Developer's His- Pacific resolution's His- band's resolution's L. Pacific resolution (Ed. Streetwood)") 11, "Mong careful" 12-10.</li> </ul>
Tent, Peter, annelli, L., Nolation, plant's HL, Norman J. Land, and S. Santang's Norman J. Particular, and plant, in surface, "In Proceedings of International Conference International Conference on Control International Conference on Conference on Conference International Conference on Conference on Conference International Conference on Conference on Conference International Conference on	Proc. Pres. North 1, Norther and No. 201, Northern M. J. Annual (Nation, Adapting American) Control (Nation), Adapting American Resolution (Nation), Nation (Nation), Nation	Proto, Deep, Strath, L., Norther, and Y. M. Norther, P. Land, J. Marker, A. Lander, and Y. Constant, P. Lander, C. Marker, Antonio P. P. Antonio and A. S. Santari, Strather, and A. S. Santari, Antonio and A. S. Santari, Strather, S. Santari, Antonio and A. S. Santari, Strather, S. Santari, S. Santari, Antonio and S. Santari, Strather, S. Santari, S. Santari, Antonio and S. Santari, S. Santari, S. Santari, S. Santari, Antonio and S. Santari, S. Santari, S. Santari, S. Santari, S. Santari, Antonio and S. Santari, S. Santari, S. Santari, S. Santari, S. Santari, Santari, S. Santari, S. Santar	Prove constrainty": 1, Notating provid- tion, "Section of the Constrainty": Notating proves provide a section of the Section of Constrainty of the Constrainty Section of Constrainty, "Section of Constrainty, Section Provide Constrainty, Section of Constrainty, Section Provide Constrainty, Section of Constrainty, Sect	201. "Weige count": L. H.". This meet heart is "promoup on the bay, "Residence" is "Street, "Annual county," and annual "Street, "Annual county," in the "Street, "International County," Inter- mentation (Internation," Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- tional County, "Inter- tional County, "Inter- tional County," Inter- Inter Inter- Inter- Inter- Inter- Inter- Inter-	The second secon	Proc. Basel, and P. L. Norman, cond- trate, Norman, S. S. Norman, cond- Norman, characteristic statements, P. Procedule of Statements and P. Procedule of Statements, Norman, S. Procedule of Statements, Norman, S.	<ul> <li>Ten, Seel, and Y. L. Norman, and Y. S. Norman, and Y. L. Norman, and S. S. Norman, and S. S. Norman, and S. S.</li></ul>	The second secon	<ul> <li>Terrer, Barris, Santo, S., Santone, and S. S. Santone, and S. Santone, and S. Santone, and S. Santone, "A second se</li></ul>		Press, Devil, Hearth, S., Millaker, Johnson, M., Waldard and S. S. Millard, Martin Joseph Jonatharyth, Nacional Jonatics, Statistics, International Metallitation, Markow, NY, Market, Market, Market, M. Market, Market, Market, M. Market, Market, Market, M.	<ul> <li>Peris, beig ment?s: E, "Ecological and the second se</li></ul>
Constanting of the second seco	Constraint, South Street, Street, St. Street, St. St. Street, St. St. Street, St.	Constant C. Street Street Street Constant of Constant Constant Con- tension of Constant Constant Con- tension Constant Constant Con- Street Constant Constant Con- Street Constant Constant Con- Street Constant Constant Con-	<ul> <li>Conserved A., Songar S., Sandara Y., P. Sanata and S. Sanata and S. Sanata Songar Songar Son</li></ul>	Construction of the second sec	Commission of the second	Company, Angel Strength, St. Strength, St. St. Strength, St. Strength, St. St. St. St. St. St. Strength, St. St. St. St. St. St.	Comparison of the second secon	Comment A., and A. Manthay I. Comment A., and A. Manthay I. Commentation of A. Martin and A. Martin, A. Martin and A. Martin, A. Martin, A. Martin, Particle and Antonio Methods and A. Particle and A. Martin, A. Martin, J. Particle and A. Martin, Martin, A. Martin, Martin, Martin, Martin, Martin, M. Martin, Martin, Martin, Martin, M. Martin, Martin, Martin	Comparing Stranger, Strenderser, Comparing Stranger, Strenderser, Comparing Strends, Strenderser, Strends, Strender, Strenderser, Strenderser, Strends, Strender, Strenderser, Strenderser, Strends, Strender, Strenderser, Strenderser, Strends, Strenderser, Strenderser, Strenderser, Strends, Strenderser, Strenderser, Strenderser, Strenderser, Strenderser, Strenderser, Strenderser, Strenderser, Strenderser, Strenderser, S		Characteristic service interaction Characteristic service interaction Characteristic Statistics (Characteristics Characteristics (Characteristics)) Marcel Statistics (Characteristics) Marcel Statistics) Marcel Statistics Marcel Statisti	<ul> <li>Conserved, Annual Street and Street St</li></ul>
Second Street St	Veneration Construction Construction Veneration and and the Construction States are a second construction product and apply the first 123, 0828	Program P., Name percent 1, 1, 1, Name percent and 12, 17 h memory and adding 17 h memory restances of 1, 10, 100 12, 1602	Property Processing and the second se	1. "Neight and" body "states", "Second States", "Second States", Luty, "Second States, States, States, States, "Second States, Stat	Transfer, New Jacobi, 10, New Jacobi, 2000, 10, 10 Lan are Sended, 10, parts (10,00, 201), 10,000 21, 274	Compared Y, New Jackson F, Karl New Jackson March M, Charl Sampler March M, Charl March M, Santa M, Charl M, Charl M, Charl	Compared Constitution of the second s	Property And Andrew Constraints of the Property of the Andrew Constraints of the second states upply of the ST 476	Company Annual Control of Control Company Annual Control of Control State and Annual Control of Control State of Control		Compared P., Mail, and M.T. D.A. Compared annual P. Comp Sciences and edited in the menty state spectra and states of the spectra and states of the spectra and spectra and spectra and spectra and spectra and spectra and spectra and	Semanular, Assan Jackson, U.S., Samanus, Assandar S. (C.A.) Samanus, Sandardh, 2018 Samanus, Sandaryan, Depth and Mill.
	-	-	Tel.	- Warner	3	E	The second	115	-		3	
wither 14	feriter 11	A PROPERTY OF	menters 17	Anator II	Sector 13	Sector 20	Heriter II.	Territor 22	a star 23	heraber 24	Carlot and a second	August 25
An anne 1997 An anne 1997 Martin anne 1997 Martin Allan Martin Allan Martin Allan	International 2015-Dis- ternational 2015-Dis- ternational and annual ment biological from biol	International Control - Co	Tenti ante Lette de la construction tente de la constru- litation de la constru- litation de la construction de la construction la const	International 2012-02- International 2012-02- International 2012-02- International 2012-02- Distance of the International Content of the International Conten	International In	Les antes en antes entre la contrata mantes la contrata des el la contrata la contrata de la contrata la contrata de la contrata la contrata de la contrata de la contrata la contrata de la contrata de la contrata de la contrata de la contrata de la contrata de la contrata de la contrata de la contrata de la contrata de la contrata de la contrata de la contrata de	The second secon	Annual and an	Las' annesis 200-12- tras annesis 200-12- tras annesis New 2-9010120 Lasta annesis 1, Marcanal Lasta annesis 1, Marcanal	An address to a second	Last automa 2014-12- transmission 2014-12- ments 2 Million Mark 2004000 1 Mark 2004000 1 Mark 2004000 1 Mark 201400 1 Mark 2014000 1 Mark 201400 1 Mark 2014000 1 Mark 2014000 1 Mark 2014000 1 Mark 2014000	Anno 1998 Anno 1998 - 1999 Anno 1998 - 1999 March 1997 Anno 1998 Anno 1998 Anno 1998 Anno 1998
Lastri, "Comp. examine start": No. "Next According to the analysis of the start of the start press, started with the start spectra start", 201, Sect. Spectra, start's 201, Sect. Spectra, start's 201, "Support."	Nami's, "Simon second a control and, "Second and Comparis" (1), Simon provide a control of the Second and a control of the Second and the control of the Second and the control of the Second and the control of the Second Second Second and the control of the Second Seco	Total", "Grand second at star", 2011, "Social Compart", "C Social Social Compart", "C Social Social Compart", Sci. Social Social Compart, Sci. Social Social Compart, Sci., Sci.	<ul> <li>Conserve Statistics (2014)</li> <li>Conserve Statistics (2014)</li></ul>	Torrill, "Search recording start", the search and the search of the search record of the search of the Search records of the search of the Search of the search of the search of the Search of the Search of the search of the Search of the Sea	<ul> <li>Construction of the second seco</li></ul>	Particle Control and Control a	Proof Contract American American Proof American American American Proof American American Proof American American Proof American American Proof American	Property, "Grant second as start") and, "searched compare", "An "Wang particular to the "wang	Presily, "Grant security dist"; http://www.presilence.com//risk/ P	Private differences and the second se	Transmission and an and a second seco	<ul> <li>Construction and a start of the start of the start of the start of the</li></ul>
Anno (L. C. and Y. Yu) Sandi (L. C. and Y. Yu) Sandi (L. C. Anno (L. C. Anno (L. C. Yu) H. Miney (L. C. Anno (L. C. Yu) Sandi (L. C. Sando) (L. Sando) (L. Sando) Sando) (L. Sando) (L. San	Annual Control of the	Anny St., St., and St. et al., Control and St. St., St. St. St., St. St., St., St.	<ul> <li>State Control State Sta</li></ul>	And A. Carrier of States o	Conception of the second secon	<ul> <li>Start (19, 10, 2017)</li> <li>Start (19, 10, 2017)</li> <li>Start (19, 10, 2017)</li> <li>Start (19, 10, 2017)</li> <li>Start (19, 2017)</li> <li>Start (</li></ul>	The provide strength of the second strength o	Annual An	Park provide and the second se	<ul> <li>Construction Construction Construction</li> <li>Construction Construction Construction</li> <li>Construction Construction</li> <li>Construction Construction</li> <li>Construction Construction</li> <li>Construction Construction</li> <li>Construction</li> <li>Const</li></ul>	<ul> <li>Constrainty, Spanning, Constrainty, Nature, A., Statel, S. S., Nature, S. Statel, S. S. Nature, Constraint, Statements, N., Nature, South, C. Malan, and N., Nature, South, C. Malan, and N. Sang, South, C. Sang, South, S. Sang, South, South, S. Sang, Sou</li></ul>	<ul> <li>Anni Jan, S., Anni Y., Ma, Manife, Standards, L., Manife, Standards, L., Manife, Standards, M., Namelander, M., Malang, Stander, Letter, Name, Standards, C., Namelander, Standards, Standard, Standard, Standards, Standard, St</li></ul>
Income priories and an interaction of the second se	Construction of the second sec	Contract of the second	<ul> <li>Second Provide Lange of the Control <pre>Second Second Se</pre></li></ul>	Contraction of the second seco	Contract of the second	Person, Paleta, Joseph Rosenson, P. Person, C. Competence and State State Person of the state of the state person of the state of the state of the state Person of the state of the state of the state Person of the state of the state of the state of the state Person of the state of the	Provide Article and the second	<ul> <li>Contact, Jackson, Lander, M., Sander, S., Sander, San</li></ul>	Construction of the second sec	<ul> <li>Territory, Particle and State (Construction), Proceeding, Particle and State (Construction), Proceeding, S</li></ul>	1. C. Manifer and C. S. S. Strandson, Astronomy N. S. Strandson, Market and S. S. Strandson, Market and S.	Partial (Marka Jorgan America)**- Press, A. Standarder, english metricles (Marka*, **) Press, parts, State**, **) Press, parts, State**, **() Press, parts,
percentage of A. Alexandro, and an article in- band, communication of A. Alexandro, A. Alexandro, March, and A. Hu, "March Sciences," March, and A. Mark, March 19, Alexandro, March, and A. Mark, Mark 19, Alexandro, March, and A. Mark, and A. Mark, and A. Mark, Mark 1997, Annual Annual A. Mark, and A. Mark, and A. Mark, and a Mark 1997, and a second A. Mark, and A. Mark, and A. Mark, Mark 1997, and a second A. Mark, and A. Mark, and A. Mark, Mark 1997, and a second A. Mark, and A. Mark, and A. Mark, and and a Mark 1997, and a second A. Mark, and and a Mark 1997, and a second A. Mark, and A. Mark, and and a Mark 1997, and a second A. Mark, and A. Mark, and A. Mark, and a Mark 1997, and a second A. Mark, and a second A. Mark, and a Mark 1997, and a second A. Mark, and a second A. Mark, and a Mark 1997, and a second A. Mark 1997, and a second A. Mark 1997, and a Mark 1997, and a second A. Mark 1997, and a second A. Mark 1997, and a Mark 1997, and a second A. Mark 1997, and a second A. Mark 1997, and a Mark 1997, and a second A. Mark 1997, and a second A. Mark 1997, and a Mark 1997, and a second A. Mark 1997, and a second A. Mark 1997, and a Mark 1997, and a second A. Mark 1997, and a second A. Mark 1997, and a Mark 1997, and a second A. Mark 1997, and a second A. Mark 1997, and a Mark 1997, and a second A. Mark 1997, and a second A. Mark 1997, and a Mark 1997, and a second A. Mark	Annotation of Carlos and Annotation 5. Second stands in the Carlos States of Annotation 5. Second states of the Carlos of Annotation 5. Second states of the Carlos of Annotation 5. Second states and the Carlos	Annual and the second s	Second se	Annual and the second s	And a second sec	Service of C. No. and Address S. Service and Association Sci. Control on the Association Sci. Control on Sci. Sci. Sci. Sci. Sci. Sci. Sci. Sci. Sci. Sci. Sci. Sci.	<ul> <li>anterestication, San and States, P. S.</li> <li>Neuroscience, Constructions, States, Science, S.</li> <li>Neuroscience, Science, Science, S.</li> <li>Neuroscience, Science, Science, S.</li> <li>Neuroscience, Science, Science,</li></ul>	Secolari and S. Ne, and Secolari S. S. Nerol, and chemistry S. Standardski, Phys. Rev. Lett. 1, 1990; 2010; 2011; 2010; 2010; 2010; 2010; 2010; 2011; 2011; 2010; 200	Anne control description is been appendix of the second state of the second state of the second state between the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state of the second state	Americanity, Nair and American Providential Control of Con- National American Strength Control Control and Control of Control Control of Control of Control of Control of Control Control of Control of Control of Control of Control of Control Control of Control of Contr	Benefiker (eff, Norgans, Steath (E. Verset, Steath (E.S., Norger, L. Jake) (1), "Steath (E.S., Norger, L. Jake) (1), "Steath (E.S., Norger, L. Jake) (1), "Steath (E.S., Steath) (1), "Steath (E.S., Steath, Steath, St. 1), "Steath (E.S., Steath, St. 1), "Steath (E.S., Steath), "Steath (E.S., Steath, St. 1), "Steath, Steath, St. 1), "Steath (E.S., Steath, Steath, St. 1), "Steath, Steath, Stath, St. 1), "Steath, Steath, St. 1), "Steath, Steath, Steath, St. 1), "Steath, Steath, Steath, St. 1), "Steath, St. 1), "Steath, St. 1), "Steath, Stath, St. 1), "Steath, St. 1), "Steath, Steath, St. 1), "Steath, St. 1), "Steath, Stath, St. 1), "Steath, Stath,	American Conf., Nac. 2021, classify 5, 5, Provid, crossin, Wooseful (M. Y. S., Pranti, Conf. (M. Yang) (C. S. Sanda), Prantisk, angle (C. Serle, Standard), Prantisk, angle (C. Serle), Standard), Prantisk, angle (C. Serle), Prantisk, a
and the last strength	13. dta	11.450	18.400	12. CB4	TT. 443	and the set of the	-	11.291	-	15. 484	No. 145	TEL SUS
erikar 17 heri ven heri sene heri seletti venen	worther: 18'	worker 29 more cert. Lan anterent introdu- lettablech attable	worker 38 hore eres has antistate repairs pression of second	serier R series motories	segiar 32 new search annals ann search annals	eccient 19 manufacture (manufac- intradient) (manufac- intradient) (manufac-	ecckar 34 marchine meters are spectrum meters are spectrum meters	eorikar 35 mar ante anter ante generate anterite tratación et prices	worker 34 notes and the antiference and the LODO-Divide Automatic	socker 37 not one second	worker 38 norm over lan weeken overette http://www.aken	worker 39 torst start. 1011/06/9/14489
And the second s	The PERSON AND A CONTRACT OF A	The Million of the State of the	period to della reast dell'Attitude (1. National et al. 2016) "Productional and and the statement of the Production of Statement of the Statement o	The Article Con- tract of the Content of the Conten	Providence of the second secon	And the construction is a second to be the construction of the proof of the construction of the the construction, the cons	The second secon	And Advantage In Advantage I Special (Carthy In) Theory, Sality and Advantage (Carthy In) Theory, Sality and Advantage (Carthy In) States, Sality and Sality (Carthy In) Sality (Car	And the second s	And the second second 1. Speed with the second sec	The second street 1. Next starts to Next sta	Period In 1970 Law Managing We Provide Construction Provide Construction of the Construction
and densities starts talk, an average to be and assess sources on supports and to an average to the average of a start to the start of average to the	And Annaly	Anna Anna Anna Anna Anna Anna Anna Anna	Neuropeopling (reprinted), Neuropeopling (reprin	head position starts with the sector ( 10) that sector ( 10) that a sector ( 10) that a sector ( 10) that a sector ( 10) that a sector ( 10)	Next Accessing Starts Mile Next Accession Starts Mile Next Accession Starts Starts Next Accession Starts Next	<ul> <li>Name (2016), and the P. (1, 1), Name (2017), Proc. (2017), Proc. Name (2017), Proc. (2017), Proc. (2017), Name (2017), Proc. (2017), Proc. (2017), Name (2017), Name (2017), Name (2017), Name (2017), Name (2017), Name (2017), Name (2017), N</li></ul>	New Constant Acade 1985 New America's State New America's State New America's State New America State	New products and the second se	Presidential score with State second roles of the State second roles of the State of the second roles of the State of the State of the second roles of the State of the State of the second roles of the S	President scale and New Access 1, 16, New Access	President and a second state of the State Association of the second state of the State of the second state of the second state State of the second state of the second state State of the second state of the State of State of the second state State of State of the second state of the second state of the State of State of the second state of the second state of the State of State of the second state of the	Performing sort in, Performing sort in, Performance and in the second Part of the second in the second Part of the second in the second Part of the second in the second in the second in the second in the Part of the second in the seco
(Comparing the second secon	Processing and participation of the second o	Provide international and international Pro- static States and Pro- tation and Distances in the second states and the Rest of the second states and the second states and the Provide Constant	Person, compared and constant's bit. "Second constant, 1, 1995, "enc. parallelout", 5, "Science, const USA, "Second of the "Second constant, edited of the "Second Constrainty, official and "Second Constrainty, 1997, Sangar, "Second Constrainty, 1997, Sangar,	Provide and a second se	Control and States and American States and	Partie Antonio S. S. Sonno P. Serte Antonio S. S. Sonno P. S. Milar, 2007 1-00, Ten Statutor 1, 5 Statutor (1977) 10, Network 1, 5 Statutor (1977)	Control and State State of the State St	Control Andrewson, and Antonio C. C. Marca and C. S. Ba, and Janes, Joney & Statemark, 1998 (C. Namarka, 2008) C. Statemark, 1998 (C. Namarka, 2008) Control and Statemark Product of the splitter of comparison of the splitter of comparison.	Preside characteristic and constraint's the Western constraint's the second second second second second second Preside second second second second second second second Preside second seco	<ol> <li>Marca and A. D. D.</li> <li>Marca and Y. C.D.</li> <li>Marca and J. D. M. (1997)</li> <li>Marca and J. D. M. (1997)</li> <li>Marca and A. M. (199</li></ol>	Construction of the construction of the second second second second second second second s	<ul> <li>Version, elevente para desentadore</li> <li>Version, elevente el terretadore</li> <li>Version, Anada, Joneto, el Secondore, acadore</li> <li>Version, elevente de la constanta de la</li></ul>
And Anne (1997) 75, and an order of the state of the feedback of the state of the state instate of the state of the state mathematics of the state of the state state of the state of the state of the state state of the state of the state of the state state of the state of the state of the state of the state state of the state of the state of the state of the state state of the state of the state of the state of the state of the state state of the state of the st	Contraction (Contraction) Contraction (Contract	Party Armit (Scott) - "- Constanting (Scott) - Controls Constanting (Scott) - Controls - Constanting (Scott) - Controls - Constanting (Scott) - Controls - Constanting (Scott) - Controls - Constanting (Scott) - Controls - Control (Scott) - Control (Scott) - Controls - Control (Scott) - Control (Scott) - Controls - Control (Scott) - Contro	Constraints (2004) 75, Constraints (2004) 1, constraints Constraints (2004) 1, constraints (2004) Constraints (2004) 1, constraints (2004)	Contraction (Contraction) Contraction (Contraction)	Contraction Contraction	Provide charge's Connect and the Connection Connect and the Connect and Connect and the Connect and Connect and the Connect and Connect and the Connect and Connect and the Connect and Connect and Connect and Connect and Connect and Connect and Connect and Connect and Connect and Connect and Connect and Connect Connect and Connect and Connect and Connect and Connect Connect and Connect and Connect and Connect and Connect Connect and Connect and Connect and Connect and Connect and Connect Connect and Connect and Connec	Print and Joseff 1997 President, Staff 1997 President, Staff 1997 President in Joseff 1997 President of Staff 1997 President o	Contraction (Contraction) Contraction (Contr	Constanting Starf 1	Profit posts from ", ", Planting, Star", 1, starting, Provide the second start ", start, with posts in second st manufactory, ", so the second st manufactory, ", so the second st Provide start start, so the second st Provide start start, so the second st Provide start start, so the second st Provide start st Provide st Prov	Paral and Alary, "A Adams of the second states, Consequent, and the second Adams of the second states and the second states and the second states and the second states and the second states and the second states and the second states and the	Part and Josef 1997 W. Particular Josef 1997 B. Particular Josef 1997 B. Particular Josef 1997 B. Particular Josef 1997 B. Particular J. Statistical Contents 1997 B. Particular J. Statistical Physics B. Particular Physics B. Particu
Holman and The Second Control (Second Control (Second Control Heat and Association (Second Control Heat and Association (Second Control Heat (Second Control (Second Control (Second Control Heat (Second Control (Second Control (Second Control (Second Control Heat (Second Control (Sec	Percent and the State of Lands Annual gard and Annual and the Pengener, Seed and the tag for showing the method and and the State of the State of the seed 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	Production (197, 1997), 1, 1997 1, National (1977), 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,	Construction on Construction Construction and period to the Construction and period to the Construction and additional period and additional period and additional period and additional period and additional period and additional period additional period and a period additional period additional period and a period additional period additional period additional period period additional period additional period additional period period additional period additional period additional period additional period additional period additional	Comparison (197), "Series Comparison Comparison (197), "Series Comparison France for electronic data from for electronic data from for electronic data from for electronic data (197)	Control and Contro	<ul> <li>Newson and State (1999). Comparison of the comparison</li></ul>	<ul> <li>Statistical (19), "Sept. Laboration Sciences, "Sent proceedings," Press, and account (19)," The process and account (19), and a sentence (19), and account only and (19), and account only account only (19), and account only (19), a</li></ul>	<ul> <li>Construction of the Construction of the Construction</li></ul>	Personal and Control Personal Accounts in Statistics and Control Personal Personal Accounts and Control Personal Personal Accounts International Accounts International Personal Per	Control and Control and Control of Control o	<ul> <li>Construction (199), Terrery, Lotter- S, Analista (2007), Construction (2007), Pressport (2007), Construction (2007), These plane advances (2007), 2007), Same plane advances (2007), particular (2007), 2007), particular (2007), particular (2007)</li></ul>	<ul> <li>Control and Control and Contr</li></ul>
	3	3	5	3	-		10.83	111	2	2	-	0.
1	miles 42.	molec di	mainer 42	autor ti	michaer 45	mainer 48.	Tana ti	Boolar II	autor D	auter 10	eleter St	mater 12
en annet mit-te- ma annet mit-te- ma 17.070 and 17.070 Annet 19.070 Annet 19.0700 Annet 19.07000 Annet 19.0700 Annet 19.0700 Annet 19.0700 Ann	International International International International International International International International International International In	International Interna- contraction (Interna- tion 4.3200) The Midd (MONS) Company of the International Company of the International Internatio	Sectore and a sector a secto	International controls word F-B200 New York Control New York Control New York Control New York Control New York Control New York Control New York Control New York Control New Y	And application (Inclusion) And application (Inclusion) Methods (Including Control of	very general mit-fre article general mit-fre werk 17,600 Nation (19,600 1 Annu (19,600) Annu (19,600) Ann	Anno 1999 Anno 1999 Anno 1997 Anno 1	Annual Control Control In Control In Control C	Very general mit-ri- contract is some work 25 35/90 Case Statement Case Contract is some	And particle con- control of the second seco	Last question to the state of a second state of the second state o	And Specific Length States of the Specific St
(a) P. Shap printing (a) by Press, and Learning (b) by press, and press (b) by press (b) and press (b) and press (b) by	<ul> <li>Control and Control and Contr</li></ul>	<ul> <li>Control, "Name and the second state".</li> <li>Control and the second state of the second</li></ul>	<ul> <li>Strail, "State straining later" <ul> <li>Strain and Later" L.</li> <li>Strain and Strain and Strain <li>Strain and Strain and Strain <li>Strain and Strain <li>Strain and Strain <li>Strain and Strain <li>Strain and Strain </li> </li></li></li></li></li></ul> </li> </ul>	Frank, "Singe second parts", in State and parts and second parts and parts and parts and here parts and parts and here parts and parts and here parts and parts and parts and parts here parts and parts and parts and parts and parts here parts and parts and parts and parts and parts here parts and parts and parts and parts and parts here parts and parts and parts and parts and parts and parts here parts and parts and parts and parts and parts and parts here parts and parts and parts and parts and parts and parts here parts and parts and parts and parts and parts and parts here parts and parts and parts and parts and parts and parts here parts and parts and parts and parts and parts and parts and parts here parts and	Prof. The court of the terms of terms	<ul> <li>Control, "These seconds of the" Inter-Second second of the Second second second of the Second second second of the Second second second second second second second Second second second second second second second second second Second second second Second second seco</li></ul>	Parity "Data access to the State Sector and Dataset" of a State Sector and Dataset" of a Sector access to the Sector access to the Sect	<ul> <li>Touris, "Destruction, const."</li> <li>Touris, and const.", "Con- transport, Annual '1 C.S., "Sensitive destruction," Con- transport, "Con- transport, "Con- transport, "Con- transport, "Con- transport," Con- Versitive const.", "Const.", "Const.", "Con- Versitive const.", "Con- V</li></ul>	<ul> <li>Control and Control of Control</li></ul>	Providenský Presez, astronom, prozefy utilis Presez, astronom, prozefy utilis Presez, prozieku astronom († 1905) Presez, prozieku astronom († 1905) Presez, prozieku astronom († 1905)	Providental Provent encodered control to the Provent encodered control to the Provent providence of the the Provent providence of the Provent providence of the the Provent providence of the Provent providence of the the Providence of the the the the the the Providence of the	ProclamB*, Propaga processory cought diff, Propaga processory cought diff, Propaga production of the state Propaga production of the state Propaga production of the state Propaga processory of the state Processory of the state Processor
Learny, Cyllicenser's Transp. Learne (L.), anner's U. Institution, anner's U. And C. Standpool and Proceedings of Management (L. Malane, Josef), Inst. Seed, South S. Malane, Josef), M. Malane, and S. S. Malane, Josef).	Property Systematic Dates, New York, States, States, States, States, States, States, States, States, States, States, States, States, States, State, New York, States, States, States, State, New York, States, States, States, State, New York, States, States, States, States, State, New York, States, State	New parts in second in the second sec	The second secon	Control Special Control (Control) Statistics of Statistics of Landon (Control) Control Statistics of Landon (Control Statistics of Landon (Control) Control Statistics of Landon (Control Statistics of Landon (Contr	Anno An Constantin State Anno An Constantin Const Constantine Constantine Constantine Const Constantine Const Constantine Const Constantine Const Constantine Const Constantine Const Cons	Territy, "generative reads," New Joint, Strawning, St. Conf., Strawning, L. Martin, Constant, L. Martin, Conf. and St. Martin, New Joint, S. Toland, and Martin, New Joint, S. Toland, and S. Toland, and Martin, New Joint, S. Toland, and S. Toland, and Martin, New Joint, S. Toland, S.	Presery, "Adversary," (2007) Presery, concerning, bio- Preside and and the Preside and the Concerning of the Concerning and Adversary Concerning and Concerning The Adversary Concerning The Adversary Concerning The Adversary Concerning Concerning and Concerning Concerning Concerning and Concerning Concerning and Concerning Concer	Trans. Parliance (* 1998) Trans. Parlies and the State Trans. Construction of the State State (* 1998) State (* 1998) Sta	Party Science (1999) Party Science (1999)	<ul> <li>Strate, Mole, Line, Y. Hu, "Support Exactly, "International Guide," Internet, Sciences, 19, 2016.</li> <li>Strate, Construction, and Internet Phys. Rev. D, Neuropean, and Internet Phys. Rev. D, Neuropean, 2016.</li> <li>Strate, Science, S. J. Hu, "Strategies, Science, Sci. 4, Science, 2017.</li> </ul>	Preserved and a server of the programmer Grand, "processory Grand," Print, and the desired State Preserved Statements of Statements Preserved Statements, and streamstoffs Etc., Neurosci. (Statements), and streamstoffs.	<ul> <li>Press, Golds, 2017; 101; "Pageson?"</li> <li>Pater, and V. Scherk, C. S. Scherk, J. S. Scherk, J. S. Scherk, S. Sch</li></ul>
And in the second desired by the second by the second seco	Construction of the Constr	Protect of the course in strength	Parties (data, reput, data and the first sector of the sec	Contraction of the second seco	Print Constitution Sector	Prove class user discoutly's Provide to be added	Probat place paper, Scentrophy Probat (whether the stage	Printer, Malan Japan, Grandard's Printer, Village Strategy	Periodi, Status, Julian, Antonios's Constitution of a graduation of a state	ALL Parties of a later descent of	No. Water of a lower of	ALL AND AND ADDRESS AD

#### Justin Johnson

Lecture 11 - 43

### Track ratio of weight update / weight magnitude

```
# 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 ~le-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

#### Justin Johnson



### 1.One time setup

Activation functions, data preprocessing, weight initialization, regularization

### 2. Training dynamics

Learning rate schedules;

hyperparameter optimization

### **3.After training**

Model ensembles, transfer learning,

large-batch training

Justin Johnson

### **Model Ensembles**

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

## Enjoy 2% extra performance

## 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!

#### Justin Johnson

#### Lecture 11 - 47

### 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)



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

Karras et al, "Progressive Growing of GANs for Improved Quality, Stability, and Variation", ICLR 2018 Brock et al, "Large Scale GAN Training for High Fidelity Natural Image Synthesis", ICLR 2019

Justin Johnson



# Transfer Learning

Justin Johnson

Lecture 11 - 49

### Transfer Learning

## "You need a lot of a data if you want to train/use CNNs"

Justin Johnson

Lecture 11 - 50

### Transfer Learning



Justin Johnson

Lecture 11 - 51

1. Train on Imagenet

FC-1000	
FC-4096	
FC-4096	
MaxPool	
Conv-512	
Conv-512	
MaxPool	
Conv-512	
Conv-512	
MaxPool	
Conv-256	
Conv-256	
MaxPool	
Conv-128	
Conv-128	
MaxPool	
Conv-64	ĺ

Conv-64

Image

2. Use CNN as a feature extractor



Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014

#### Justin Johnson



1. Train on Imagenet

FC-1000 FC-4096 FC-4096 **MaxPool** Conv-512 Conv-512 MaxPool Conv-512 Conv-512 MaxPool Conv-256 Conv-256 MaxPool Conv-128 Conv-128 MaxPool

Conv-64

Conv-64

Image



Conv-512

MaxPool

Conv-512

Conv-512

MaxPool

Conv-256

Conv-256

MaxPool

Conv-128

**Conv-128** 

MaxPool

Conv-64

Conv-64

Image

**Classification on Caltech-101** 



Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014

October 9, 2019

#### Justin Johnson

1. Train on Imagenet

FC-1000 FC-4096 FC-4096 MaxPool Conv-512 Conv-512 MaxPool Conv-512 Conv-512 MaxPool Conv-256 Conv-256 MaxPool Conv-128 Conv-128 MaxPool Conv-64

Conv-64

Image

t 2. Use CNN as a

Image

feature extractor



Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014

#### Justin Johnson

Lecture 11 - 54

1. Train on Imagenet

FC-1000 FC-4096 FC-4096 MaxPool Conv-512 Conv-512 MaxPool Conv-512 Conv-512 MaxPool Conv-256 Conv-256 MaxPool Conv-128 Conv-128 MaxPool Conv-64

Conv-64

Image

2. Use CNN as a

Image

feature extractor



Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014

#### Justin Johnson

Lecture 11 - 55

### 1. Train on Imagenet

FC-1000

2. Use CNN as a feature extractor

FC-4096

FC-4096

MaxPool

Conv-512 Conv-512

MaxPool

Conv-512

Conv-512

MaxPool

Conv-256

Conv-256

MaxPool

Conv-128

Conv-128

MaxPool Conv-64

Conv-64

Image

### Image Classification



Razavian et al, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", CVPR Workshops 2014

#### Justin Johnson

#### Lecture 11 - 56

### 1. Train on Imagenet

2. Use CNN as a

FC-4096

FC-4096

MaxPool

Conv-512

Conv-512

MaxPool

Conv-512

Conv-512

MaxPool

Conv-256

Conv-256

**MaxPool** 

Conv-128

Conv-128

MaxPool

Conv-64

Conv-64

Image

feature extractor



Justin Johnson





Razavian et al, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", CVPR Workshops 2014

#### Lecture 11 - 57

#### Transfer Learning with CNNs 3. Bigger dataset: 2. Use CNN as a 1. Train on Imagenet **Fine-Tuning** feature extractor FC-1000 Continue training FC-4096 FC-4096 FC-4096 Remove CNN for new task! FC-4096 FC-4096 FC-4096 last layer MaxPool MaxPool MaxPool Conv-512 Conv-512 Conv-512 Conv-512 Conv-512 Conv-512 MaxPool MaxPool MaxPool Conv-512 Conv-512 Conv-512 Conv-512 Conv-512 Conv-512 MaxPool MaxPool MaxPool Freeze Conv-256 Conv-256 Conv-256 these Conv-256 Conv-256 Conv-256 MaxPool **MaxPool** MaxPool Conv-128 Conv-128 Conv-128 Conv-128 Conv-128 Conv-128 MaxPool MaxPool MaxPool Conv-64 Conv-64 Conv-64 Conv-64 Conv-64 Conv-64 Image Image Image

#### Justin Johnson

Lecture 11 - 58

1. Train on Imagenet

FC-1000
FC-4096
FC-4096
MaxPool
Conv-512
Conv-512
MaxPool
Conv-512
Conv-512
MaxPool
Conv-256
Conv-256
MaxPool
Conv-128
Conv-128
MaxPool
Conv-64

Conv-64

Image

2.	Use CNN	N a	as a
	FC-4096 FC-4096 MaxPool Conv-512 Conv-512 MaxPool Conv-512		Remove last layer
	Conv-512 MaxPool Conv-256 Conv-256 MaxPool Conv-128 MaxPool Conv-128 MaxPool Conv-64 Conv-64		Freeze these

# 3. Bigger dataset: **Fine-Tuning**

FC-4096	-
FC-4096	
MaxPool	
Conv-512	
Conv-512	
MaxPool	
Conv-512	
Conv-512	
MaxPool	
Conv-256	
Conv-256	
MaxPool	
Conv-128	
Conv-128	
MaxPool	
Conv-64	
Conv-64	
Image	

Continue training CNN for new task!

### Some tricks:

- Train with feature extraction first before fine-tuning
- Lower the learning rate: use ~1/10 of LR used in original training
- Sometimes freeze lower layers to save computation

#### Justin Johnson

Lecture 11 - 59



Lecture 11 - 60

### Transfer Learning with CNNs: Architecture Matters!

### ImageNet Classification Challenge



Improvements in CNN architectures lead to improvements in many downstream tasks thanks to transfer learning!

#### Justin Johnson

#### Lecture 11 - 61

### Transfer Learning with CNNs: Architecture Matters!

**Object Detection on COCO** 



Ross Girshick, "The Generalized R-CNN Framework for Object Detection", ICCV 2017 Tutorial on Instance-Level Visual Recognition

#### Justin Johnson

#### Lecture 11 - 62



#### Justin Johnson

Lecture 11 - 63



#### Justin Johnson

![](_page_64_Figure_1.jpeg)

#### Justin Johnson

![](_page_65_Figure_1.jpeg)

#### Justin Johnson

![](_page_66_Picture_1.jpeg)

![](_page_66_Figure_2.jpeg)

Karpathy and Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions", CVPR 2015

Girshick, "Fast R-CNN", ICCV 2015 Figure copyright Ross Girshick, 2015. Reproduced with permission.

#### Justin Johnson

Lecture 11 - 67

![](_page_67_Picture_1.jpeg)

Karpathy and Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions", CVPR 2015

Girshick, "Fast R-CNN", ICCV 2015 Figure copyright Ross Girshick, 2015. Reproduced with permission.

Justin Johnson

Lecture 11 - 68

![](_page_68_Figure_1.jpeg)

![](_page_69_Figure_1.jpeg)

- 1. Train CNN on ImageNet
- 2. Fine-Tune (1) for object detection on Visual Genome
- 3. Train BERT language model on lots of text
- 4. Combine (2) and (3), train for joint image / language modeling
- 5. Fine-tune (5) for image captioning, visual question answering, etc.

Zhou et al, "Unified Vision-Language Pre-Training for Image Captioning and VQA", arXiv 2019

#### Justin Johnson

#### Lecture 11 - 70

Transfer learning is pervasive! Some very recent results have questioned it

**COCO** object detection

![](_page_70_Figure_2.jpeg)

Training from scratch can work as well as pretraining on ImageNet!

... If you train for 3x as long

He et al, "Rethinking ImageNet Pre-Training", ICCV 2019

October 9, 2019

Justin Johnson

Transfer learning is pervasive! Some very recent results have questioned it

COCO object detection

![](_page_71_Figure_2.jpeg)

Pretraining + Finetuning beats training from scratch when dataset size is very small

Collecting more data is more effective than pretraining

He et al, "Rethinking ImageNet Pre-Training", ICCV 2019

October 9, 2019

#### Justin Johnson
Transfer learning is pervasive! Some very recent results have questioned it



COCO object detection

My current view on transfer learning:

- Pretrain+finetune makes your training faster, so practically very useful
- Training from scratch works well once you have enough data
- Lots of work left to be done

He et al, "Rethinking ImageNet Pre-Training", ICCV 2019

#### Justin Johnson

Lecture 11 - 73

# Distributed Training

Justin Johnson

Lecture 11 - 74

### Beyond individual devices





<u>Cloud TPU v2</u> 180 TFLOPs 64 GB HBM memory \$4.50 / hour (free on Colab!)

<u>Cloud TPU v2 Pod</u> 64 TPU-v2 11.5 PFLOPs \$384 / hour

#### Justin Johnson

Lecture 11 - 75



This image is in the public domain

Justin Johnson

#### Lecture 11 - 76

Idea #1: Run different layers on different GPUs



This image is in the public domain

Justin Johnson

Lecture 11 - 77

Idea #1: Run different layers on different GPUs Problem: GPUs spend lots of time waiting



This image is in the public domain

Justin Johnson

Lecture 11 - 78

Idea #2: Run parallel branches of model on different GPUs



This image is in the public domain

Justin Johnson

Lecture 11 - 79

Idea #2: Run parallel branches of model on different GPUs Problem: Synchronizing across GPUs is expensive; Need to communicate **activations** and **grad activations** 



This image is in the public domain

Justin Johnson

Lecture 11 - 80





Justin Johnson

Lecture 11 - 81

Forward: compute loss



Justin Johnson

Lecture 11 - 82

Forward: compute loss



**Backward: Compute gradient** 

Justin Johnson

Lecture 11 - 83

Forward: compute loss



Justin Johnson

Lecture 11 - 84



#### Justin Johnson

Lecture 11 - 85

### Mixed Model + Data Parallelism



Example: https://devblogs.nvidia.com/training-bert-with-gpus/

Justin Johnson

Lecture 11 - 86

### Mixed Model + Data Parallelism



Example: https://devblogs.nvidia.com/training-bert-with-gpus/

Justin Johnson

Lecture 11 - 87

Justin Johnson

Lecture 11 - 88

# Suppose we can train a good model with one GPU





#### How to scale up to dataparallel training on K GPUs?

















 $\bullet$   $\bullet$   $\bullet$ 

#### Justin Johnson

#### Lecture 11 - 89

# Suppose we can train a good model with one GPU





**Goal**: Train for same number of epochs, but use larger minibatches. We want model to train K times faster!

### How to scale up to dataparallel training on K GPUs?

















• • •

#### Justin Johnson

Lecture 11 - 90

# Large-Batch Training: Scale Learning Rates

Single-GPU model: batch size N, learning rate α





Alex Krizhevsky, "One weird trick for parallelizing convolutional neural networks", arXiv 2014 Goyal et al, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour", arXiv 2017 K-GPU model: batch size KN, learning rate Kα

















#### Justin Johnson

#### Lecture 11 - 91

### Large-Batch Training: Learning Rate Warmup



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

October 9, 2019

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

#### Justin Johnson

Lecture 11 - 92

### Large-Batch Training: Other Concerns

#### Be careful with weight decay and momentum, and data shuffling

For Batch Normalization, only normalize within a GPU

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

	lust	in	hn	SO	n
2	ust.			30	

Lecture 11 - 93



## Large-Batch Training: ImageNet in One Hour!



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

#### Justin Johnson

#### Lecture 11 - 94

Goyal et al, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour", 2017 Batch size: 8192; 256 P100 GPUs; 1 hour



Goyal et al, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour", 2017 Batch size: 8192; 256 P100 GPUs; 1 hour

Codreanu et al, "Achieving deep learning training in less than 40 minutes on imagenet-1k", 2017 Batch size: 12288; 768 Knight's Landing devices; 39 minutes

Goyal et al, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour", 2017 Batch size: 8192; 256 P100 GPUs; 1 hour

Codreanu et al, "Achieving deep learning training in less than 40 minutes on imagenet-1k", 2017 Batch size: 12288; 768 Knight's Landing devices; 39 minutes

You et al, "ImageNet training in minutes", 2017 Batch size: 16000; 1600 Xeon CPUs; 31 minutes

Goyal et al, "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour", 2017 Batch size: 8192; 256 P100 GPUs; 1 hour

Codreanu et al, "Achieving deep learning training in less than 40 minutes on imagenet-1k", 2017 Batch size: 12288; 768 Knight's Landing devices; 39 minutes

You et al, "ImageNet training in minutes", 2017 Batch size: 16000; 1600 Xeon CPUs; 31 minutes

Akiba et al, "Extremely Large Minibatch SGD: Training ResNet-50 on ImageNet in 15 Minutes", 2017 Batch size: 32768; 1024 P100 GPUs; 15 minutes Recap

### 1. One time setup

Activation functions, data preprocessing, weight initialization, regularization

### **2. Training dynamics**

Learning rate schedules;

hyperparameter optimization

### 3. After training

Model ensembles, transfer learning,

large-batch training

**Last Time** 

Today

Lecture 10 - 99

# Next Time: Recurrent Neural Networks

Justin Johnson

Lecture 11 - 100