ROB 498/599: Deep Learning for Robot Perception (DeepRob)

Lecture 10: Training Neural Networks - Part 2 02/12/2025





Today

- Feedback and Recap (5min)
- Training NNs
 - Learning Rate scheduling
 - Choosing Hyperparameters
 - Model Ensembles, Transfer Learning
- Summary and Takeaways (5min)



Recap

1. One time setup:

Last time

• Activation functions, data preprocessing, weight initialization, regularization

2. Training dynamics:

Today

 Learning rate schedules; large-batch training; hyperparameter optimization

3. After training:

• Model ensembles, transfer learning



1. One time setup:

• Activation functions, data preprocessing, weight initialization, regularization

2. Training dynamics:

Now

• Learning rate schedules; large-batch training; hyperparameter optimization

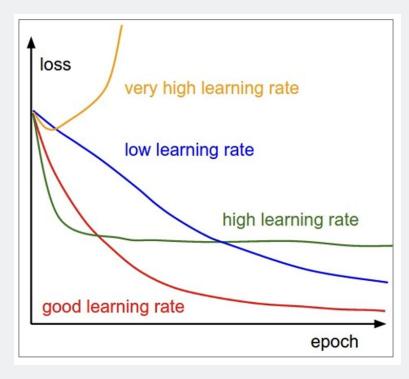
3. After training:

• Model ensembles, transfer learning



Learning Rate Scheduling

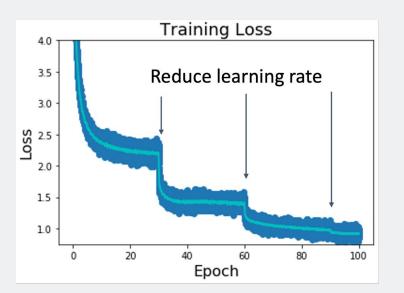
SGD, SGD+Momentum, Adagrad, RMSProp, Adam all have learning rate as hyper parameter



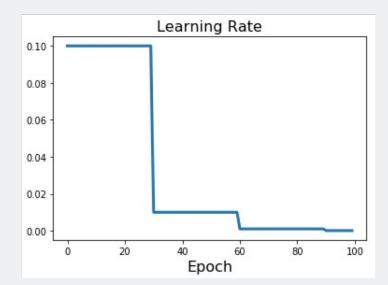
Q: Which one of these learning rates is **best** to use?



Learning Rate Decay: Step

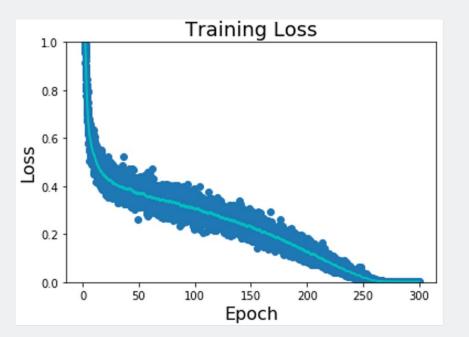


Step: Reduce learning rate at a few fixed points. E.g. for ResNets, multiply LR by 0.1 after epochs 30, 60, and 90.



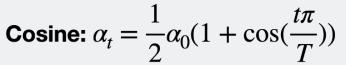


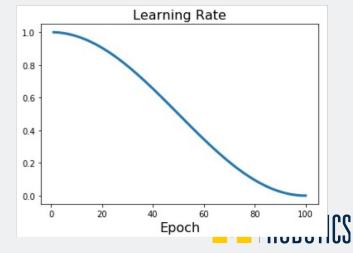
Learning Rate Decay: Cosine



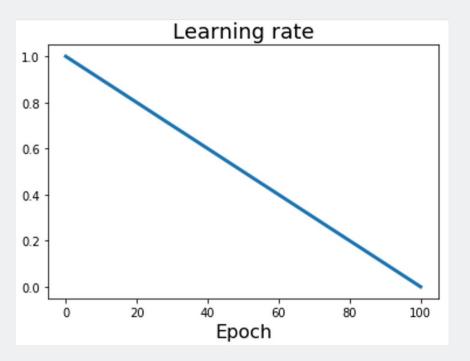
Loshchilov and Hutter, "SGDR: Stochastic Gradient Descent with Warm Restarts", ICLR 2017 Radford et al, "Improving Language Understanding by Generative Pre-Training", 2018 Feichtenhofer et al, "SlowFast Networks for Video Recognition", ICCV 2019 Radosavovic et al, "On Network Design Spaces for Visual 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.





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.

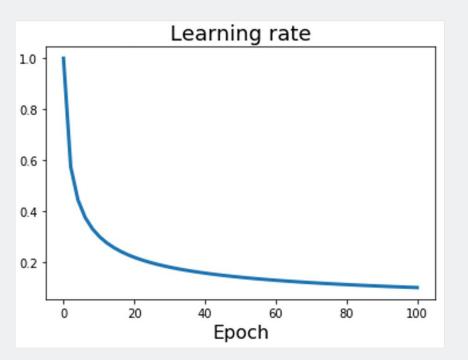
Cosine:
$$\alpha_t = \frac{1}{2}\alpha_0(1 + \cos(\frac{t\pi}{T}))$$

Linear: $\alpha_t = \alpha_0(1 - \frac{t}{T})$

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", 2019Yang et al, "XLNet: Generalized Autoregressive Pretraining for Language Understanding", NeurIPS 2019



Learning Rate Decay: Inverse Sqrt



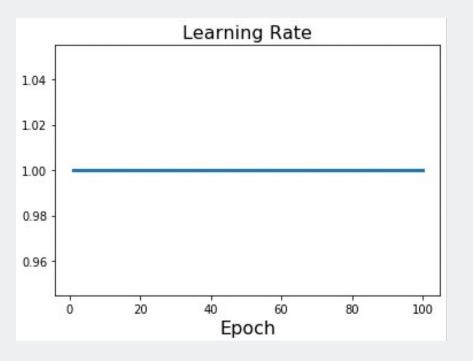
Step: Reduce learning rate at a few fixed points. E.g. for ResNets, multiply LR by 0.1 after epochs 30, 60, and 90.

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

Linear: $\alpha_t = \alpha_0(1 - \frac{t}{T})$
Inverse sqrt: $\alpha_t = \alpha_0/\sqrt{t}$



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.

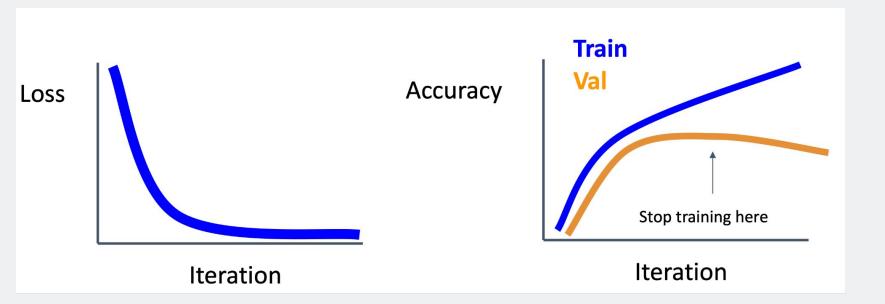
Cosine:
$$\alpha_t = \frac{1}{2}\alpha_0(1 + \cos(\frac{t\pi}{T}))$$

Linear: $\alpha_t = \alpha_0(1 - \frac{t}{T})$
Inverse sqrt: $\alpha_t = \alpha_0/\sqrt{t}$
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



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!

Choosing Hyperparameters: Grid Search

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

Example: Weight decay: [1x10⁻⁴, 1x10⁻³, 1x10⁻², 1x10⁻¹] Learning rate: [1x10⁻⁴, 1x10⁻³, 1x10⁻², 1x10⁻¹]

Evaluate all possible choices on this hyperparameter grid



Choosing Hyperparameters: Random Search

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

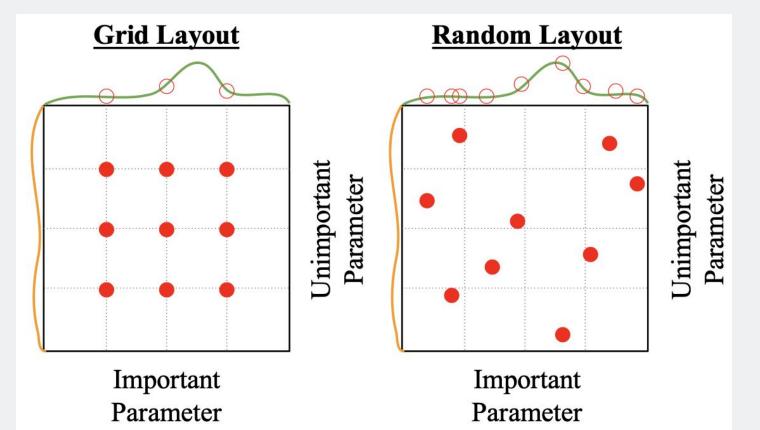
Example:

Weight decay: log-uniform on [1x10⁻⁴, 1x10⁻¹] Learning rate: log-uniform on [1x10⁻⁴, 1x10⁻¹]

Run many different trials



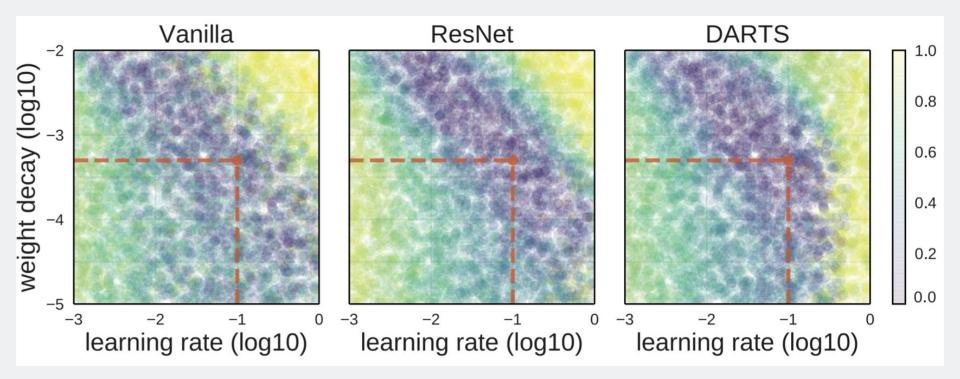
Hyperparameters: Random vs Grid Search



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



Choosing Hyperparameters: Random Search



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



(without tons of GPUs)

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 mini batches); 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 loss
- Step 2: Overfit a small sample
- Step 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 learning rates 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

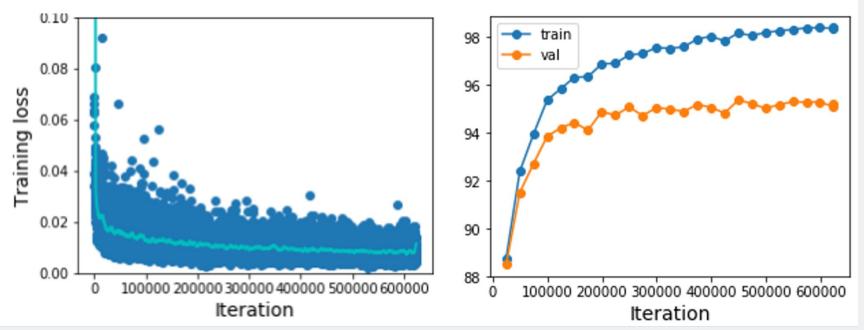
Pick best models from Step 4, train them for longer (~10-20 epochs) without learning rate decay



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

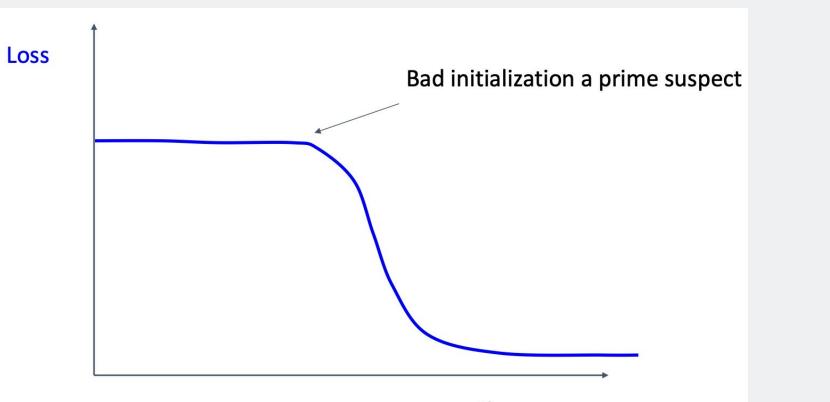


Looking at Learning Curves



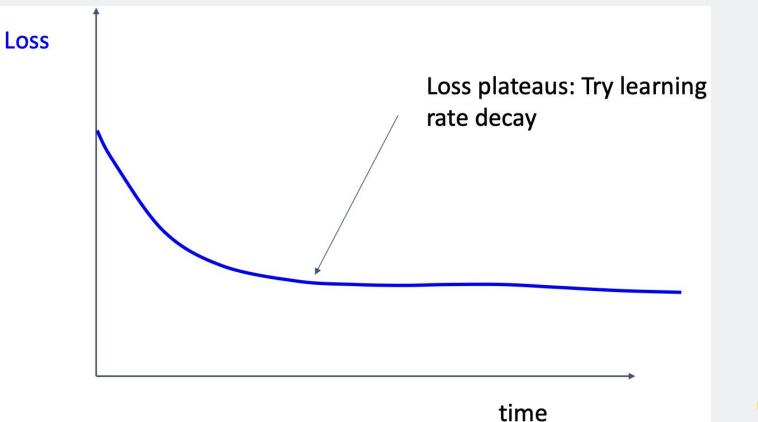
Losses may be noisy, use a scatter plot and also plot moving average to see trends better



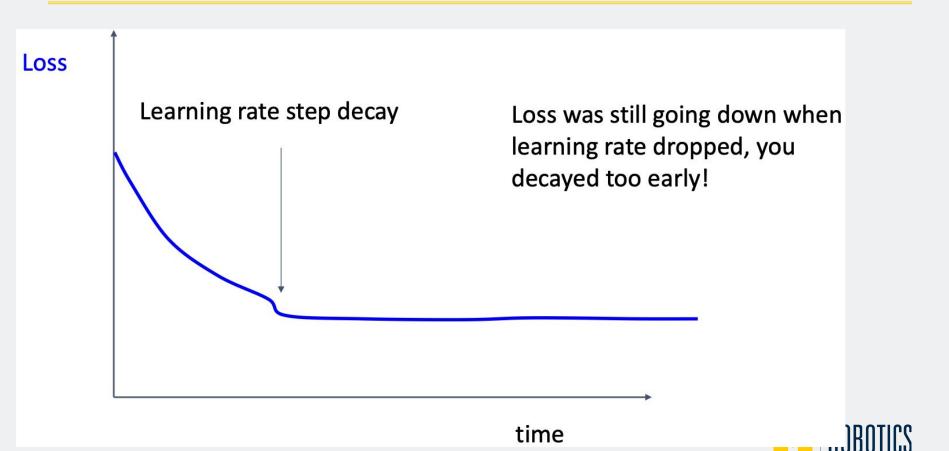


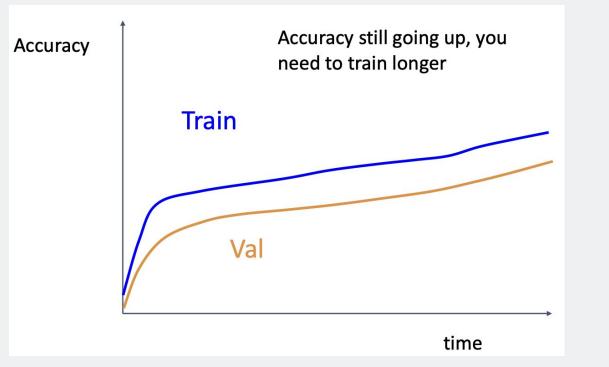




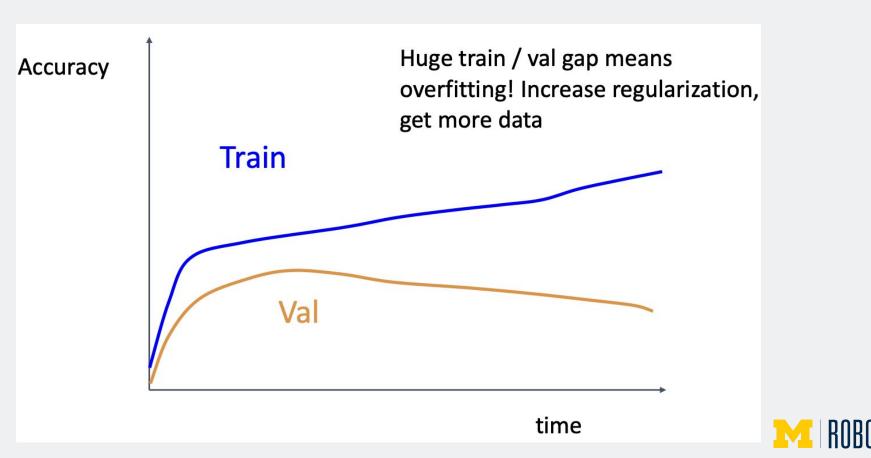


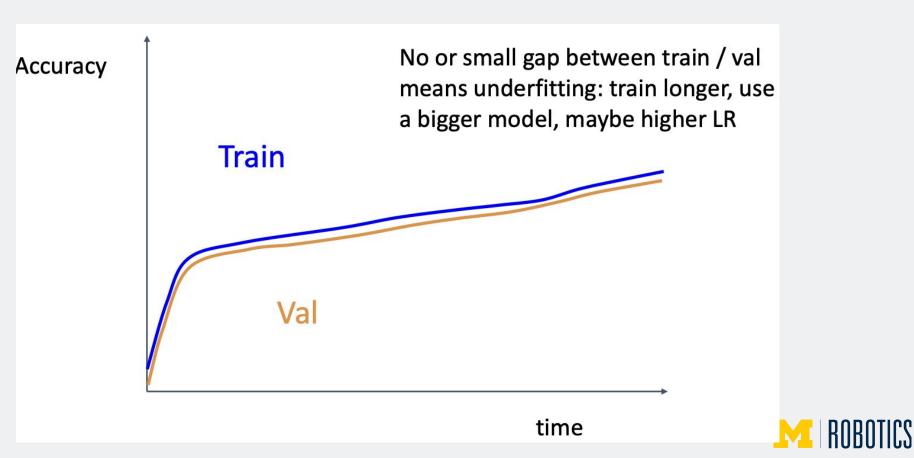
ROBOTICS











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 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



This image by Paolo Guereta is licensed under CC-BY 2.0

Cross-validation "command center"

Bonduneni (Smanne)	Marrier J	perfore 1	Decement 1	Sector 1	Jerrine 4	Dearber 7	Seater 1	accier 2	Sector 17	16	Contract 11	Decement 12
in annual printing	Anter State J inter anternal, Millionan- merce Jan 18/00 Inter State States		the spread to be			and and the part of the local		and water (1) and	and when the state		Martine 33 with and with an and the second man 1 along the difference	
- 141.00,100	Martin da 10/00	Special & Allow	Martin 20. Lo. 10	ments to 17100	State State State	And Address (Con-Co- construction of the Con- construction of the Con- truction of	And Address of the owner.	Amount 14 15/10	March 1 42/10		specific & station	and sectors in the local secto
No. of Concession, Number of Street, or other	Test, mid. Mondard's A. Parryan's	The second secon	Apple or ground of the Party of	C. April (1997) 191	And Advantages of the American's	Party of a suspensive to be being and	The set of the second of the second s	April 1000 100	April 1997 Constants of the Paris		Apple (1997) the	And And And Annual States of the Owner, or other Designation of the Owner, or other De
	The second secon		Arts, Markel And Comparis, 197	Courts and an and the second		The second and the second seco		tern, Warner, Ann, Statistical Co., 1977	And Address of the Ad		No. Name and American Street	The second secon
Apartanta and a start with	Parts Produce (1997) 105, San, Springer, St. Press, Mark, and C. St. Vingenith.	Contraction over 1915	Sector and a sector and	Constanting state in-	Contraction and in American	Rendered and the start of the	Party and a start in the second	And and a start of the second	Party and a survey of the second		The second secon	References and the
a provinsi and	And provide the second second	Print, Spinster, Spinster,	A send, "planning", fundi-	Proof Anna and I the Property	Anna an an an an an an	And Mr. In alternation	And Street, Constraint Const.	Party Street,	And Antonio and Antonio and		Contraction of Contra	And States and States
Name of Street Street	Party industrial dat. Construct?	Nuclear Statistics and Community of Statistics	And the Annual State Street P.	Conf. Statistics, pr. Manager, P.	Anna and an art and a second s	Party Statistics and Street P.	Party management and investor of	 An interaction and interaction An interaction of the plane. 	And the American State Street, or a street, or a street, and the state of the street, and the		L. Sharp reprint the	Part of Annalised Art. Street
	And Addition in the Addition of the Addition o		P.S. Stable of L.S.	The part front is front parts		ter, balandi bi berne, rama, angan, anartanti		And Addition in the American	And Address of the Ad		AL, "Anderson", 5	C. Sector S. S.
and in the second se	Anna and a start of the	Provide a constraint of a second seco	Name and Address of the	Constant of the second states	Anna and analysis	Provide the second seco	Particular and south of	Providence of the second secon	Name and Address of the second s		And Andrewson, and	And a state of the
	Construction and a second second		Conception and Conception			Contraction and A. L. S. (1971).		Contact Contact, Statistics	President Street, Street, Street, St.		Continue alors, 5 allows, Continue al contact, Street, et al.	Construction of the Armonia of Street,
	Statute of the pair start, it.		Summer of the and many to	And the second s		Annual and the set or all a large set of the		Annual of the part of the local of the second secon	Section, and the second of the		Annual of the proceeding of	And the Contract of the second states
(100 m) 10 10 m	Construction of the Constr		Council and Council Statements	Theory and the first the		Constant and look forests		Personal and a real beaution of the second s	Particle and and forest		Annual Contract of	Construction of the owner of the
	tion det admitistication		Come pare annual (1, 201) Some pare annual (1, 201) Second Come pare 14, 2011			From part barrier of the		Print and another of the	From par around the rest		total part incontants of a	Some air and the first
	12.053	11.640	13,004	117.812	13.174	14.474	24.248	147.1816	14.803		13.349	24.352
	1.1		-			-	71				- 1	-
	-		-	- Wanybries	-		-	21			-1	
				-			-					
	1	1	1	and and an other strength		N		A	n	1	-	A
enter 10	Acrise 13.	Second and Advanced and Advanced	acation 17	Angles 18		and the state	sector 25	winiter 21	Bearing 23	Access 21	annines 23	and the
	month in child	manufactor and state of the local division o	The second distance	ments at 19,100	State of the local division of the local div	manufacture and an other	speed in \$1.00	territoria anticado	Married A Laboratory	And A local division	appent of \$10,000	mark in Will
a and the second s	Transformer in Amagent	The second is the second secon	And the second s	Contraction of the American Street, Street, and Street, Street	And Address of the Ad	Appendix and a second state of the second stat	Part of the second seco	Appendix and a second s	State Property in Street of the Street of Stre	And Address of the American Street, or Stree	And Adding to Annual A	And the second second
	And Advantage of the second se		Control Andrewski, Street,			And Address and Addres		And Article State Compared in Con-	Control Control (1997) 1 (1	Construction of Street, Street, Street, Street, Married, Married, Street, Stre	Contraction of the local division of the loc	The Arrist Control of Control
And And And And And And	The second second second	Section and the sector set.	New Access of the Accession	And Annual and a start of the	Statistics and the	Providence and the second	Providence and the	Part Annual and the Annual	Paral Land, and a real former	Construction of the second state	Party and Arriston 7.7.5. Party and an arrival and a second state of the second state	Print Association and and
	The second secon		Press, Andrewson, Name,		Contraction of Contract	Contraction of the second states of the second stat		Annual Agencies Street	Parallelia di Angela da Carto	Contraction of the Arrival States	And Anna and A. St. Toward,	Start Subarts Start
The contract of the second of	Antis, resistored and research	No. 12 Constant of Articles	Nation Contacts and Contact?	Parts compared and threatery?	Annual consistence and former of	Party contacts (g) therein(d)	Partie Instantic at America's	And indiana at the P	Party control of terms of	Carl Annal and a summer	And Address of American	And the constraint of the second
And the second second	Antonio Paramo Antonio Al	and Annual and Annual Annual Pro-	P.C. Name of A. Stranger	Contraction of the second second	The Superior States and the	And Address of the American Street	And Address of Address	Arts, Mandalo M. S. Standard, Street Street	And Address of the American	And	The second second second second	Contraction in the second
and a second second second	And and the set of the	Section and and	Name and a start of	Part and read and	Annalasi ang atawa	And the second s	Prove make house in	Party and a state of	A COLOR OF THE REAL PROPERTY AND	Contraction of the second second	A read of the second se	And in the second second
	Conception and Links		Comparing story 1, 2000			Principal Street, Street, Street, Principal Street, St		Providence singly 1, Million, Principal R. (1993), Manual 1997	A statistical and an anti-		Charged St. Jorgen, Sciences 471	Control of States, States
	Antonio of the part shout it.		Second of Contract States			Providence, Science, and St.		Annual and Annual and Annual a	Nacional Contraction of the	Services of the part start, 1 Services (Services)	Annual of the sector of the	State of the local division of the local div
	Construction of the Constr		A second second second second			American's and American		Annal and state the second	A Control of the Arrival State		Annual An	Anna and a second
	Process prime annualized by the pro-		Compared and an other	Print and according to the second sec		Party protection of the		These products and a local state of the second	State Nor Apple 51, pro-		From any Social Control of the	Presid prick providents in a price
414	13.018	LL AND	1.3. 480	12.088	22.462	11.4%	14.508	11.2%		21.484	14.1.65	14.385
	-		11	211 11 11		51	110	-1	i ha	24		-11111
2		2	-	1		2					- Charles	-
-										5	1	
urban 27	Gentler 18	acceler 19	Desiliar II	Berner H	Section 32	Berline 11	Decriter 14	Decement 21	Deriver 24	Sector 11	Decema 18	December 10
And Address of the local division of the loc	and general part of	and the state	ter and second	terms with an and the second s	And In Column	out water divide	State and a state of the state	and the second second	and states in the	State of the second sec	And A DOWN	and statement includes
en pa falte e altes caltes	spinit in at/no	Inguine and an inclusion	aparte de la la la	The second second	again, ph. 02/20 prime passible building	Number of Street	spart of target	Appendix 1. Social	and an and a second sec	againt in the St	againt & Stories	Annal in Article
The Charles of American	And the state of the second	Party Print Print Print	A Design of the American Street, Stree	Provide a state of the second state of the sec	The second second second	Party of the second of the same second	April 1997 by Arriston Street,	And Street, Street, Street, Street,	And the second second second	And Alexandra a result	And Children & wanners	And Alexandra and the
	Party and Millington and		The second secon			And An American Control of State		Print, and and and a start	Contraction and the		Statistical States	The second second second
A REAL PROPERTY AND ADDRESS OF	And Annal And any sugar	Contraction in the supervised	Statut Barry Trees	Next Arrive and Arrangements	And Andrew Street, or a supervised	And Annual in Annual Inc.	Party and American Street, Str	Number of Street, Stre	And Address of the Ad	And And And - on Anna	And Roots's Linear strength	And
and an alter of the	Park and in statistic to	Party March 1, description of the	No. 40, in ann 1, No. Nach Anna 1,	New Art is directly in-	Condition in the	Annual Street, American's Proof.	Normality of second state	No. of A. Barris B.	National Action in	Not all it death the	No. of Academy Street,	And and the second second second
"Name long" 1 . The loss of the	Ph. "Mont start" (20), Prot. Sant Scott 1, Women, and	No. "Barriston" C. M.	11. "Role start V. He was not load to be seen and	And the state of t	 Yang shaft C.M. Yang Shaft C.M. Yang Shaft Sh	Antis destants in them (P)	No. "Road cane". S. 20, Store cane lines. S. Woman, conf.	The Colorence of the Co	No. State start CA.	A Constant Carl	No. Want Law? Coll.	And Street, Start, S. W. Street, of
	Parameter of the second	Part of our line town	Contraction of the distance	Contraction and American	Constant of the Assessed	A Colorado a construction	And a state of the local division of the	Tim constraint	Contraction of the summary of		Party of the second distance.	Charl Provide Assess
	Ford American State		Contraction of the local			A start in the print of the start		Number of States	North Address of the Party of t		Stort Avenue Avenue	Column and a
	Control Incontrol Station		Control Control Street and			Contraction of the		Contractory of the Contractory of the	Contraction of the		Contraction in the	Con Nume Scient
ALL DATE OF LEAST AND ADDRESS OF LEAST	Tarth and beautiful to	Read and the second second	And Alexandra And Anna	Statistical Statistics	The local state of the local sta	Real risk in fraid time	Paul and in the second second	Name and Provident Lands	Nach and the frage hits	Name and Address of August Aug	And the part of the	23,70,70,70,742
	Company, Sona property () in		Name of Array and the Array			Press of the second very life.		Constrainty of the second state of the second	A support of the second		Annual State and Control of the local state of the	Summer Street, second 1, 1, 2
And Address and a linear	Address of the start (10.775	Annald Contact and a Good St.	Concept Land and London Toronto.	the set of the sector of the	TAXABLE INCOMENTS OF TAXABLE	The set	Distantion of the second secon	11. TH	NAME AND ADDRESS OF ADDRESS OF	Total and the second se	Lot 141	all AND
								- 1 I I	21 1 1	21111	100	-1 1 1
-					-	- 44	1	21	1	2	1.00	
	1		-	1.1	-1	- Manufacture	120	21	7	1.1	1.00	
And shares and	- Constitution	1 and a second second	A CONTRACTOR OF A	SPECIFICA STR	1.	Considered and	A CONTRACTOR OF A	1 to a to read	CARLED BURN			1 August Colors
alteri 45	anahos \$1	marker 43 sector and and	anaims 43.	makes \$1	miniber 45	section 25	Bergins all	accrete an	Incluse 13 Constant of the second Constant of the second Constant of the second	mighty 54	widther Till	erarbet 10
	man of the second	much 4 32/32	New Orlean Stream	March 1 (1971)	Section 11, 14,150	Annual of All	Annual Approximation	Married 17, North	mand 14 (WW)	March 1 (19) M	And in Fight	and it. Alth
Street and the second	And and a subscription of the second	Print Manager and Street Street, Stree	A state of the second s	Construction of the second	Paral Market Sciences	Annual Statements	Contraction of the local division of the loc	State Print Collins	Contraction in the local division of the loc	And straight and	And all the second states	Contraction of the local division of the loc
	Annual Property Street		Printle, "Data second and the			Statistic Print, Science Statistics		State of the second state	And the second second second	Construction proving	Particular and a loss	And Advantage of the
and providence of the state of the	Party Contractor Contractor		Participation and the state		And American Street and Street an	Party Content of the second	Service of the servic	Party of the second sec	And States and Add.		NAME AND ADDRESS OF TAXABLE PARTY.	And Station in the
	Freider, Samer ausser, and Sameraria	Const. Const. and C. Dig. Const."	Annual in and a line of the second of the se		And Annal and An	And an in a second sta	And all is and in the	Property and the second second	And an arrange and a second se	Statute Lab. on August	Statistical and second	Si Alina -
anger begennen berefen.						and the second of the		Tarrant, Marcare, 1.	Martin Station, Sc.			and the second state of th
and finite sector from the sector of the sec	A real distants in constants	Party managers in research	And the Annual State of States of States	Particulation of an eveneration	Part of Charleson's An Dissource Pri-	and the second states	and the second second second	A. Salar sport Links	St. Anna series Cale	Pault design in the	Party destant in summary	Party of the second second
		And	And a second sec			and the second	And the second s	A share and the state of the st	And	the second second	And and a second second	The second secon

Tensorboardwandb.ai

https://docs.wandb.ai/tutorials/ pytorch/



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 ~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



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



Model Ensembles

- 1. Train multiple independent models
- 2. At test time average their results:

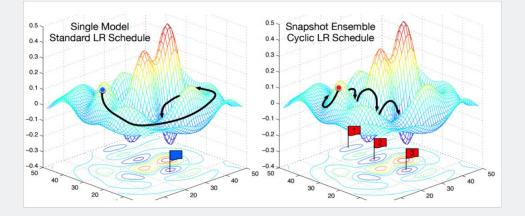
(Take average of predicted probability distributions, then choose argmax)

Enjoy 2% extra performance

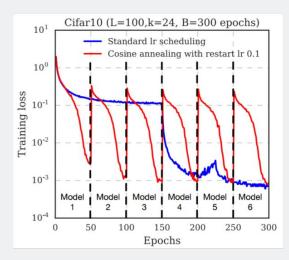


Model Ensembles: Tips and Tricks

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!



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)

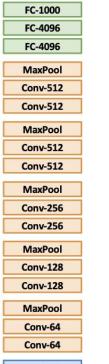
```
while True:
    data_batch = dataset.sample_data_batch()
    loss = network.forward(data_batch)
    dx = network.backward()
    x += - learning_rate * dx
    x_test = 0.995*x_test + 0.005*x # use for test set
```

Polyak and Juditsky, "Acceleration of stochastic approximation by averaging", SIAM Journal on Control and Optimization, 1992. Karras et al, "Progressive Growing of GANs for Improved Quality, Stability, and Variation", ICLR 2018Brock et al, "Large Scale GAN Training for High Fidelity Natural Image Synthesis", ICLR 201

Transfer Learning:

Generalizing to New Tasks

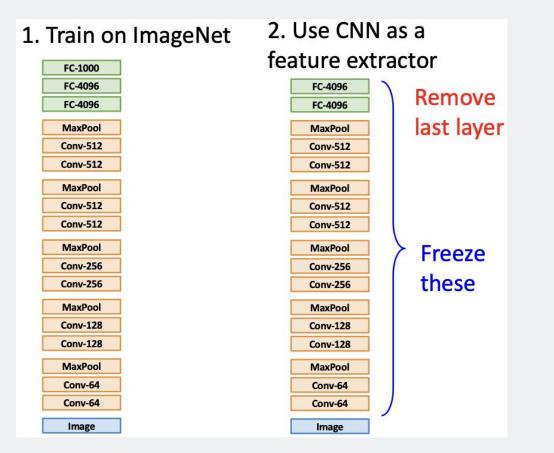
1. Train on ImageNet



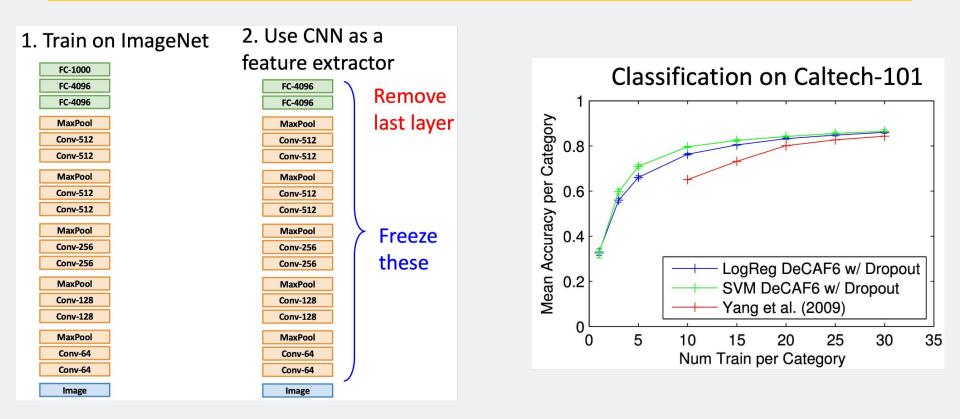
Donahue et al, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ICML 2014 https://arxiv.org/abs/1310.1531



Image

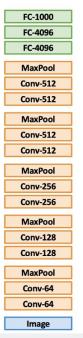




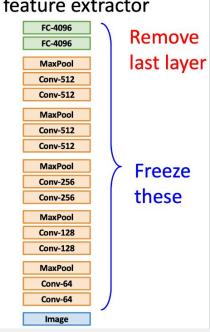


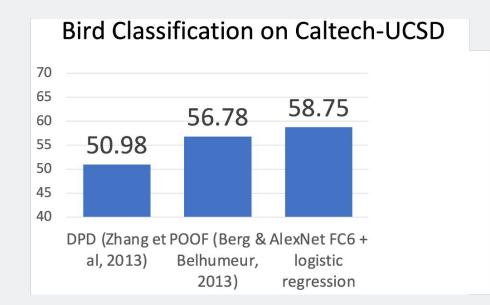


1. Train on ImageNet



2.	Use	CNN	as a	
C				

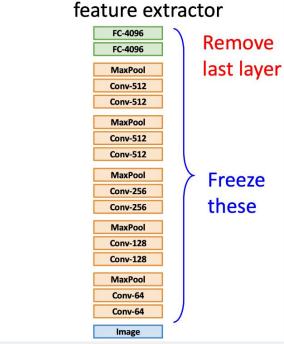




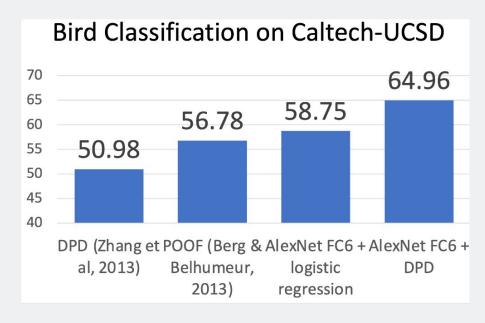


1. Train on ImageNet

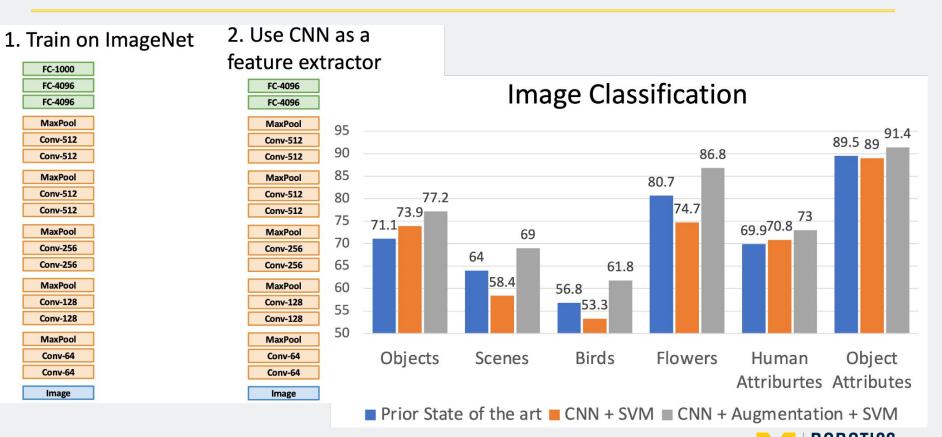
FC-1000 FC-4096 FC-4096 MaxPool Conv-512 Conv-512 MaxPool Conv-512 Conv-512 MaxPool Conv-256 Conv-256 MaxPool Conv-128 Conv-128 MaxPool Conv-64 Conv-64 Image



2. Use CNN as a

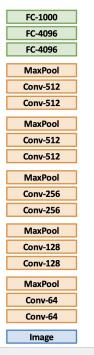




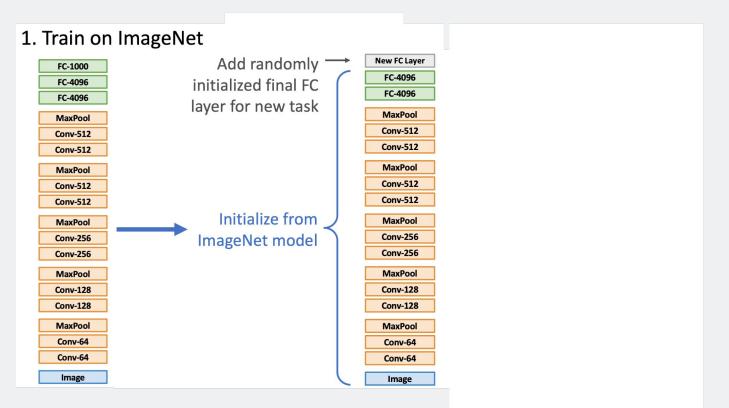


Razavian et al, "CNN Features Off-the-Shelf: An Astounding Baseline for Recognition", CVPR Workshops 2014 https://openaccess.thecvf.com/content_cvpr_workshops_2014/W15/papers/Razavian_CNN_Features_Off-the-Shelf_2014_CVPR_paper.pdf

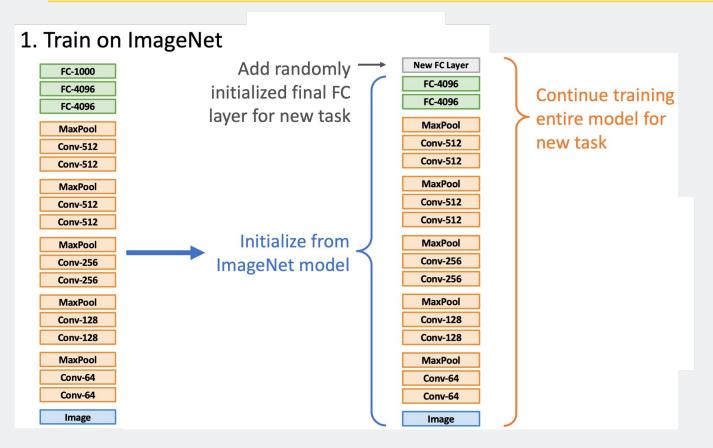
1. Train on ImageNet





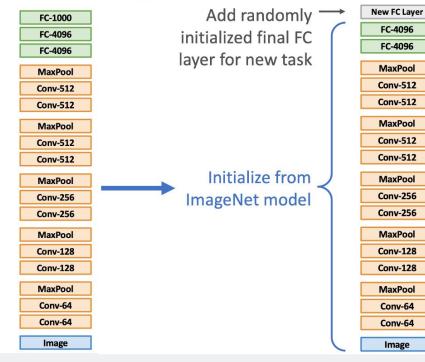








1. Train on ImageNet



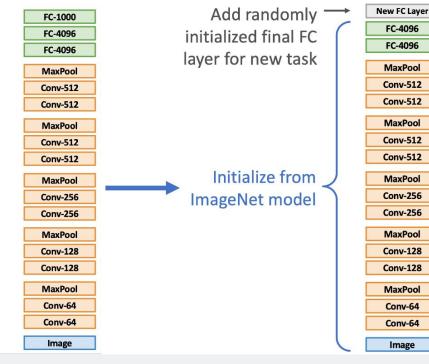
Continue training entire model for new task

Some tricks:

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



1. Train on ImageNet



Continue training entire model for new task

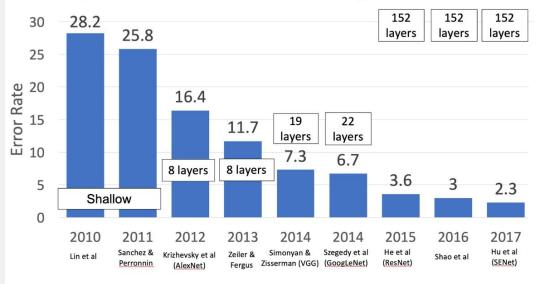
Compared with feature extraction, fine-tuning:

- Requires more data
- Is computationally expensive
- Can give higher accuracies



Transfer Learning: Architecture Matters!

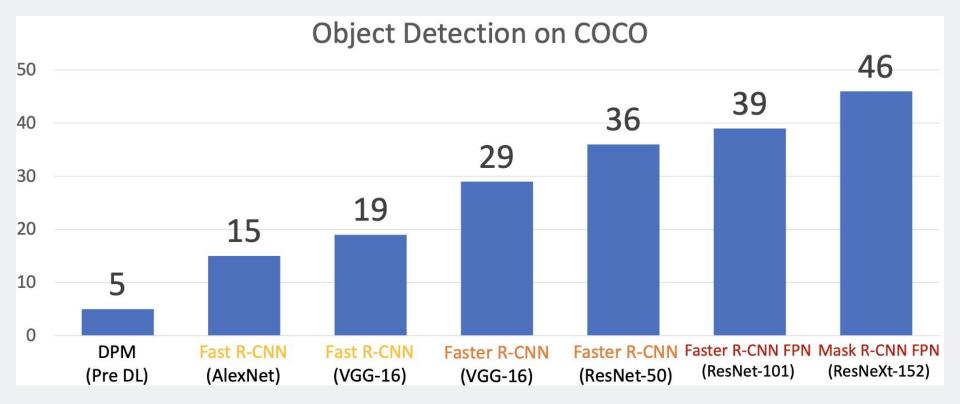
ImageNet Classification Challenge



Improvements in CNN architecture leads to improvements in many down stream tasks thanks to transfer learning!

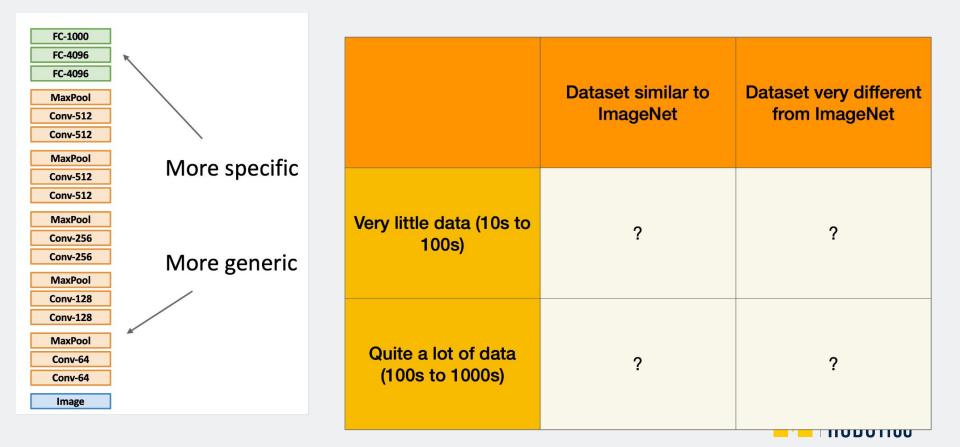


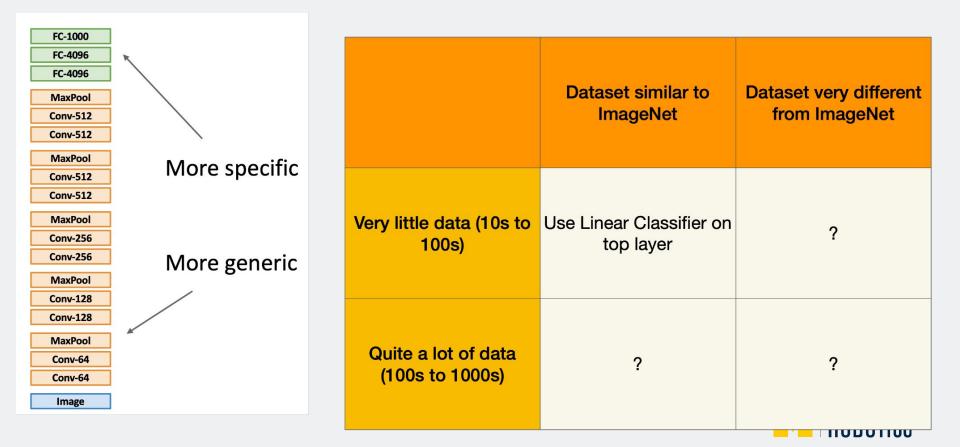
Transfer Learning: Architecture Matters!

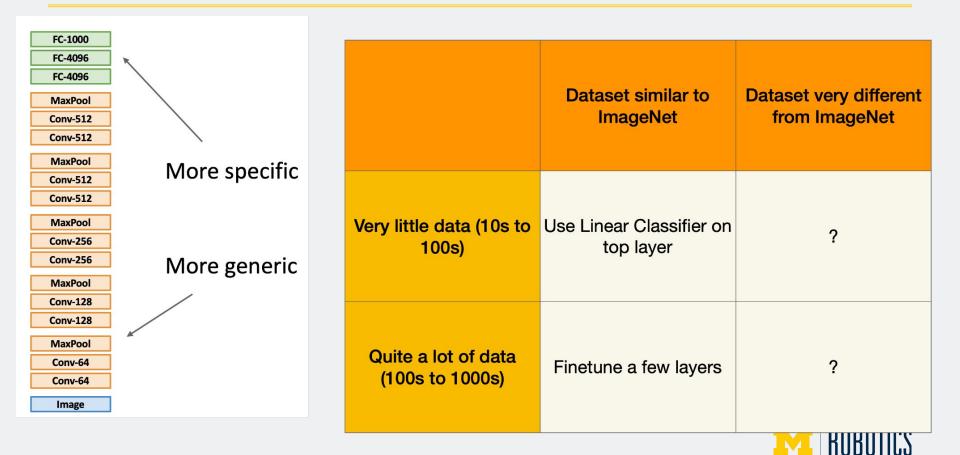


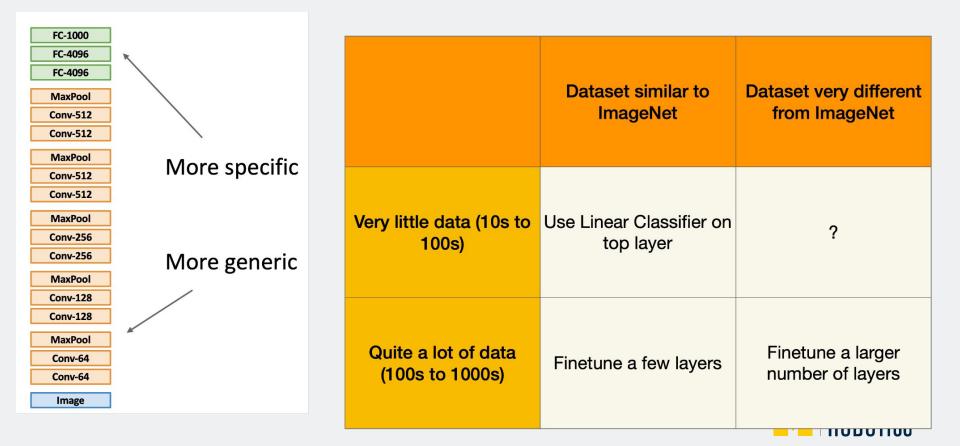
Ross Girshick, "The Generalized R-CNN Framework for Object Detection", ICCV 2017 Tutorial on Instance-Level Visual Recognition https://instancetutorial.github.io/

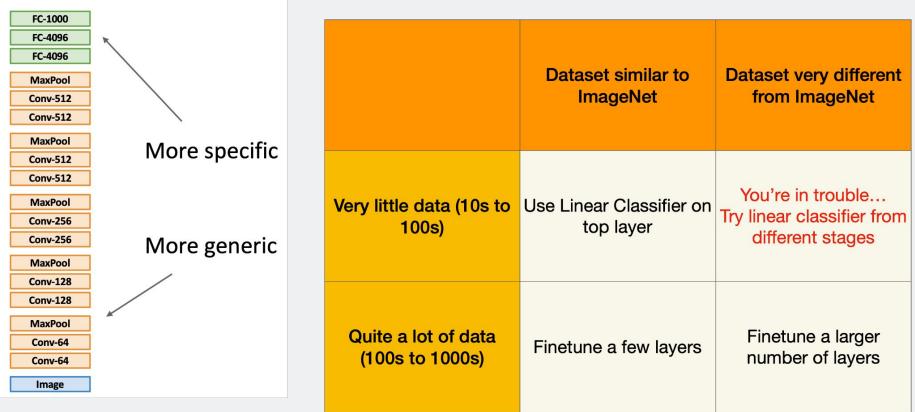




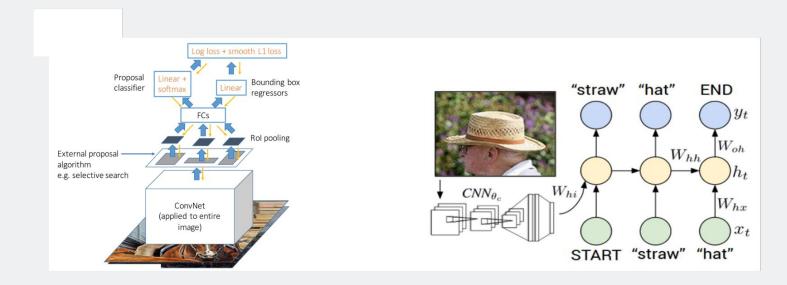








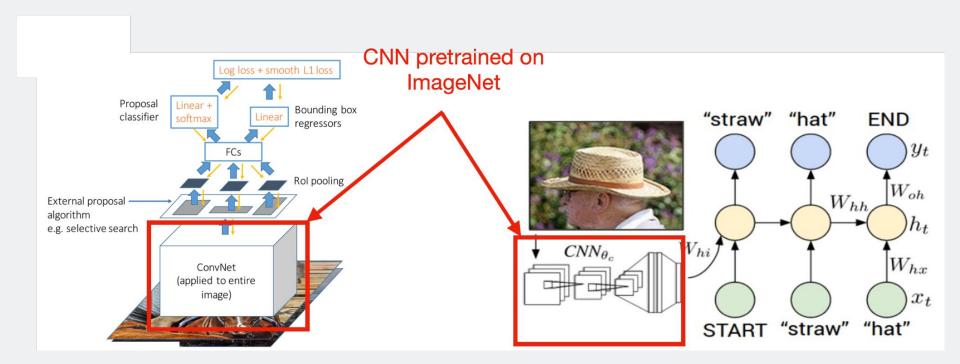




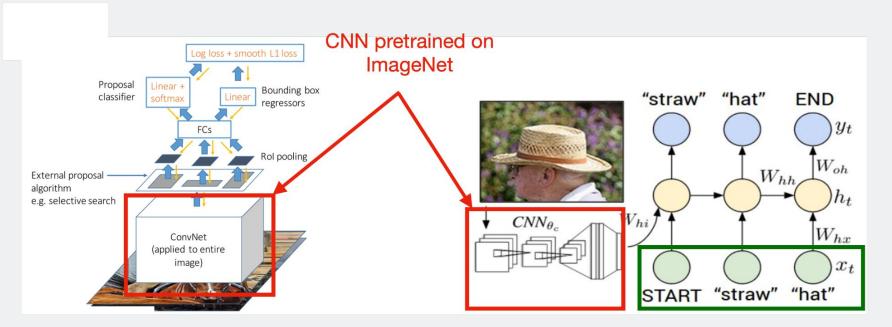
Girshick, "Fast R-CNN", ICCV 2015 Figure copyright Ross Girshick, 2015. Reproduced with permission

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



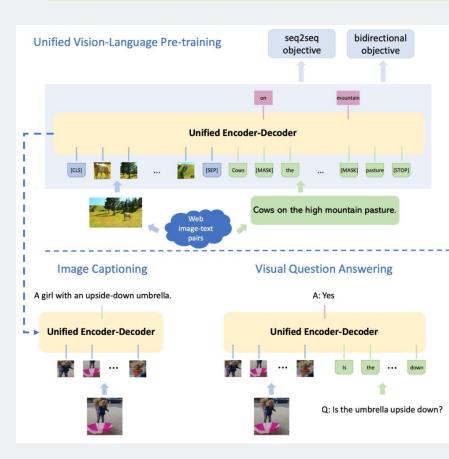






Word vectors pretrained with word2vec

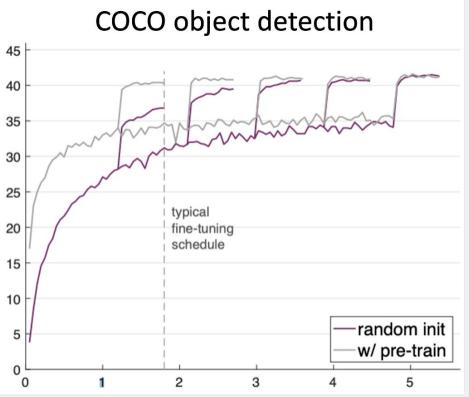




- 1. Train CNN on ImageNet
- 2. Fine-Tune (1) for object detection on Visual Genome
- 3. Train BERT language model on lots of text
- 4. Combine (2) and (3), train for joint image / language modeling
- 5. Fine-tune (5) for image captioning, visual question answering, etc.



Transfer Learning: Helps you converge faster!



If you have enough data and train for much longer, random initialization can sometimes do as well as transfer learning

He et al, "Rethinking ImageNet Pre-Training", ICCV 2019 https://openaccess.thecvf.com/content_ICCV_2019/papers/He_Rethinking_ImageNet_Pre-Training_ICCV_2019 paper.pdf



Transfer Learning: Helps you converge faster!

Pretraining for Robotics (PT4R)

Workshop at the 2023 International Conference on Robotics and Automation - ICRA London, May 29 2023, full-day workshop

Very active area of research!



Summary

1. One time setup:

- Activation functions, data preprocessing, weight initialization, regularization
- 2. Training dynamics:
 - Learning rate schedules; large-batch training; hyperparameter optimization
- 3. After training:
 - Model ensembles, transfer learning



Next time

Today