Experimental Evaluation

Ameet Talwalkar

10-718 Data Analysis Course
Feb 27, 2019
slide credit: Aarti Singh

Course Announcements

- Assignment 2 to be released tonight
- Topics for Experimental Evaluation talks will also be released tonight

Goal: Critically evaluate the quality of an empirically focused ML paper

- Choose a recent ML paper with significant empirical section to read, some examples:

  ImageNet Classification with Deep Convolutional Neural Networks
  Do CIFAR-10 classifiers generalize to CIFAR-10?

- We will provide a list of ~10 papers
  - Everyone in subgroup must present a different paper

- Submit a writeup for each presented paper in your subgroup
  - What is the general ML problem being studied in this paper?
  - What is the purpose / stated goal(s) of the empirical evaluation?
  - What are a few positive and/or negative aspects of the experimental evaluation?

Goal: Critically evaluate the quality of an empirically focused ML paper

- 15 min presentation
  - General problem (3 mins)
  - Experimental setup and results (5 mins)
  - Critique setup / results (4 min)
  - Questions / discussion (3 min)

Can be challenging to present!

Presentation tips

Summary / Experimental Setup (3 mins).

- Can't explain full scientific contributions
- Focus on high-level ideas / providing context

Present/critique the empirical setup / results (9 minutes).

- Focus first on high-level approach of the empirical evaluations
- Use your judgment to determine what details are needed to present your critique

Questions / discussion (3 minutes).

- Ask for audience opinions before stating your own critiques

Some questions to consider when critiquing experimental evaluation

- What is the purpose / stated goal(s) of the empirical evaluation?
- Is the experiment design reasonable given the stated goals?
- Are the stated goals achieved?

Case Study 1

ImageNet Classification with Deep Convolutional Neural Networks

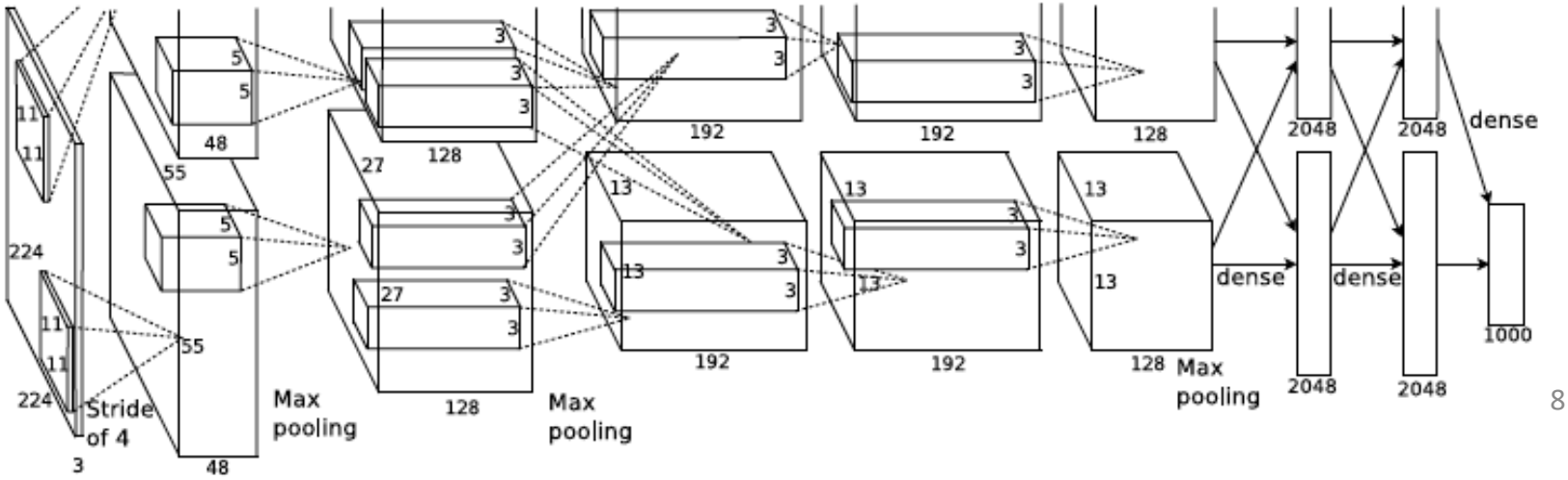
Case Study 1

ImageNet Classification with Deep Convolutional Neural Networks

Summary of paper: Largest CNN trained with highly optimized GPU implementation gives best result to date on ImageNet subsets used in ILSVRC-2010 and ILSVRC-2012 competitions

Few tricks:

- For faster training: ReLU units, local normalization, overlapping pooling
- For preventing overfitting: Data Augmentation and Dropout



Case Study 1

ImageNet Classification with Deep Convolutional Neural Networks

Experimental design:

1.2 million images, 1000 categories Task 1 Classification

- Top-1 and Top-5 test error rates on ISLVRC-10 (test labels publicly available) for their basic CNN
- Top-1 and Top-5 validation and test error rates on ISLVRC-12 (test labels NOT publicly available) for their basic CNN and some variations
- Top-1 and Top-5 test error on Fall 2009 version of ImageNet with a variation of their basic CNN

Case Study 1

ImageNet Classification with Deep Convolutional Neural Networks

Experimental results:

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

Goals achieved: Mostly, only compare 1 task and 1 metric though (ISLVRC-10 had a hierarchical metric too that was removed in ISLVRC-12)

More later…

Some questions to consider when critiquing experimental evaluation

- What is the purpose / stated goal(s) of the empirical evaluation?
- Is the experiment design reasonable given the stated goals?
- Are the stated goals achieved?
- **Are appropriate baseline methods considered?**

Baselines are extremely important: biased classes

Accuracy of classifier

	Mean
• Classifier 1	92%
• Classifier 2	87%

Training dataset had 93000 normal patients and 7000 patients with cancer

Baselines are extremely important: multiple classes

Accuracy of classifier

	Mean
• Classifier 1	52%
• Classifier 2	44%

Training dataset 10000 images: 2 classes, 5000 images each

Training dataset 10000 images: 10 classes, 1000 images each

Baselines are extremely important: regression

Accuracy of regressor

	Mean Squared Error
• Regressor 1	25
• Regressor 2	100

Standard deviation of data ~7

MSE vs $R^2 := 1 - MSE/Variance$

(Fraction of variance explained by predictor)

Baselines are extremely important: alternative approaches

Simple approaches

- Mean of data
- Unregularized estimate
- Linear predictors
- …

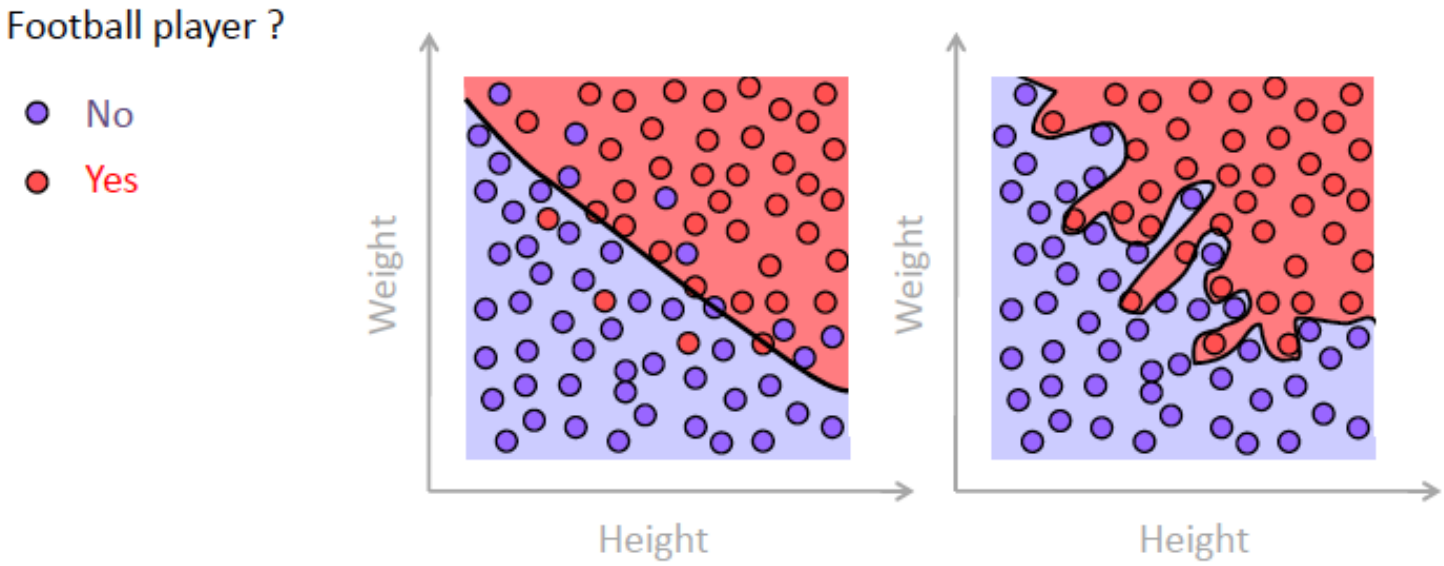
State-of-art approaches

- alternative methods in recent prior work

Some questions to consider when critiquing experimental evaluation

- What is the purpose / stated goal(s) of the empirical evaluation?
- Is the experiment design reasonable given the stated goals?
- Are the stated goals achieved?
- Are appropriate baseline methods considered?
- **Are appropriate evaluation metrics used?**

Critical to report testing and NOT training accuracy



Regression example: Training R^2 0.9 in predicting activity at one brain region using activity at another brain region Test R^2 0.01

Model fit example: Training likelihood 0.99, Testing likelihood 0.3

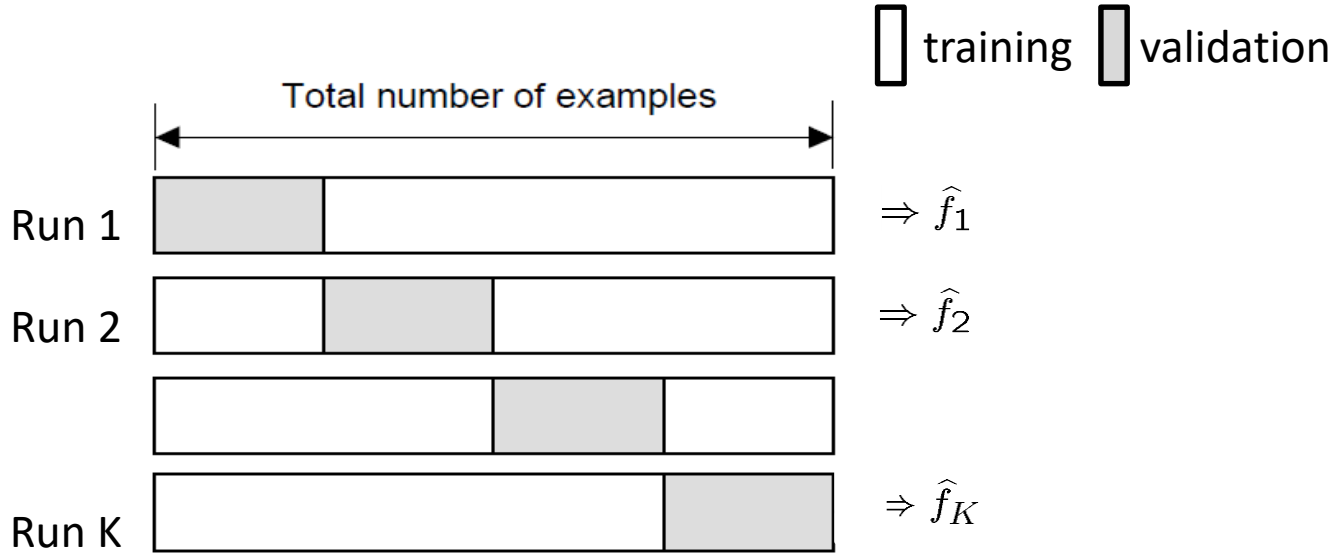
Best run test accuracy doesn't make a classifier better

Accuracy of classifier

	Average	Best run
• Classifier 1	92%	97%
• Classifier 2	87%	100%

Reusing cross-validation data will not reveal test accuracy

- Train K predictors using K-fold cross-validation



- Choose best predictor as one with largest accuracy on cross-validation data
- Report its accuracy on cross-validation data as test accuracy

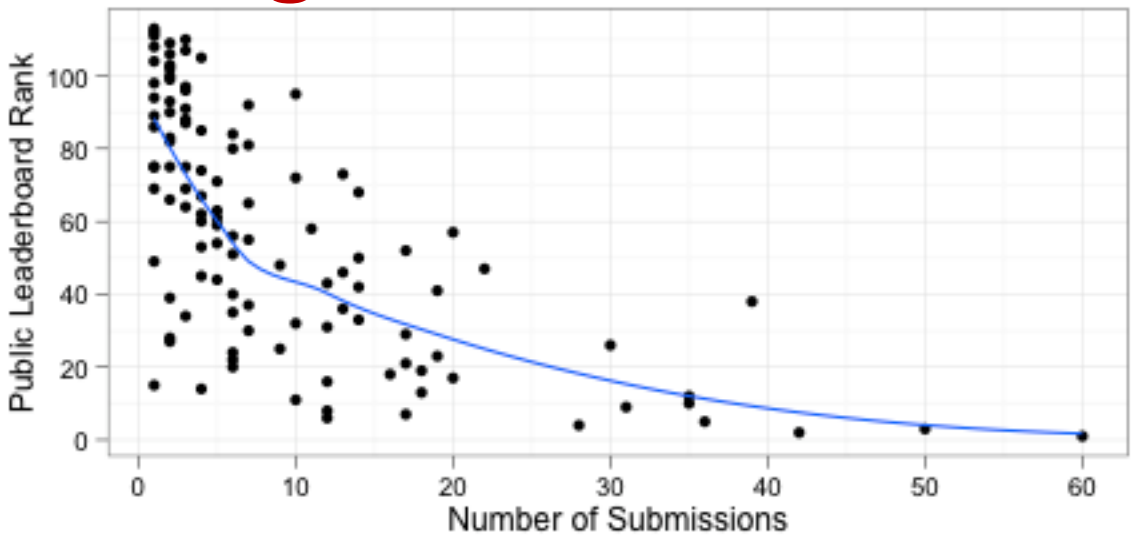
Test set accuracy may not reveal generalization performance

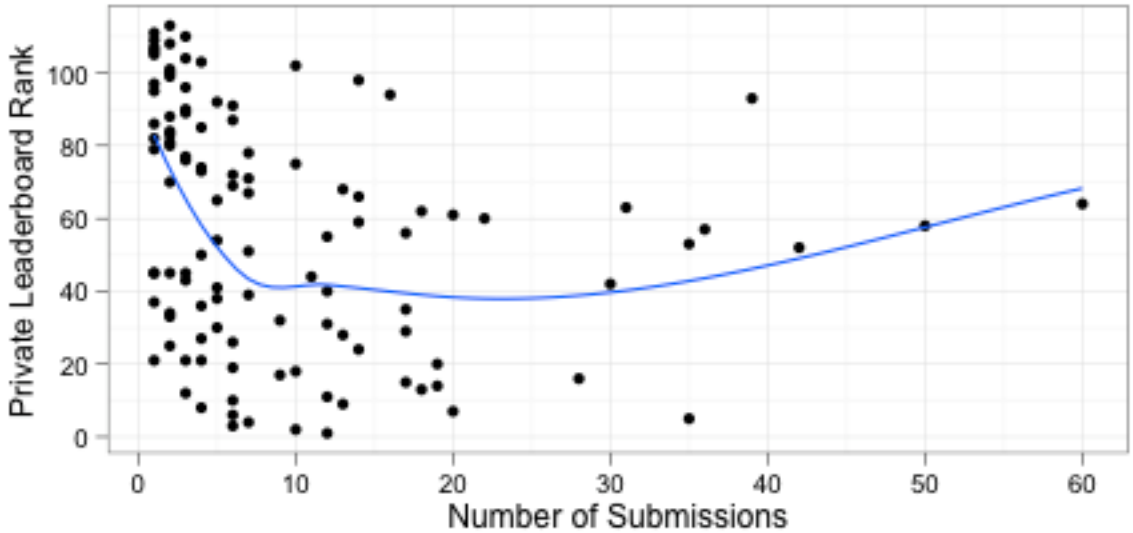
- Standard test sets often reused over and over by ML researchers – may lead to overfitting on test set
- http://gregpark.io/blog/Kaggle-Psychopathy-Postmortem/

#	Δ1w	Team Name	MCAP	Entries	Last Submission UTC (Best Submission - Last)
1	-	willkurt *	0.85295	60	Tue, 26 Jun 2012 03:41:54 (-21.9d)
2	-	Greg Park	0.85195	42	Fri, 29 Jun 2012 01:08:38 (-14.1d)
3	-	Luca Massaron	0.85163	50	Thu, 28 Jun 2012 14:51:04 (-25.6d)
4	-	Stein	0.85093	28	Thu, 28 Jun 2012 08:15:16 (-9.8d)
5	-	NoisyServer	0.85082	36	Fri, 29 Jun 2012 21:21:52 (-21.9d)

#	Δ1w	Team Name	MCAP	Entries	Last Submission UTC (Best Submission - Last)
1	-	y_tag *	0.86997	12	Tue, 26 Jun 2012 12:46:19
2	↑1	Bruce Cragin	0.86745	10	Fri, 29 Jun 2012 22:28:17 (-47.6h)
3	new	Indy Actuaries	0.86700	6	Fri, 29 Jun 2012 03:40:38 (-3.4d)
4	↓2	jontix	0.86697	7	Thu, 31 May 2012 11:05:41 (-9.9d)
5	-	JKARP	0.86683	35	Fri, 29 Jun 2012 22:02:34 (-3.3d)

Accessing test data too often can lead to overfitting





Test set accuracy may not reveal generalization performance

- Standard test sets often reused over and over by ML researchers – may lead to overfitting on test set
- Check out *The Ladder: A Reliable Leaderboard for Machine Learning Competitions* [Blum, Hardt, 2015]

Competitions use 3(4) types of data splits:

1. Training data
(split into training and cross-validation)

2. Public Test/Leaderboard data

3. Private test data

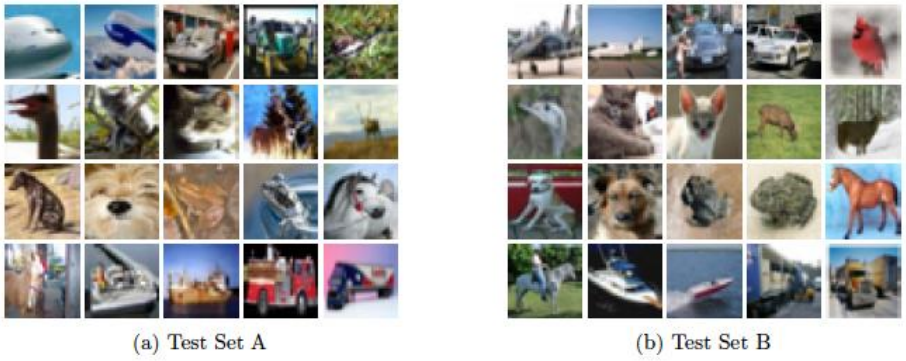
Case Study 2

Do CIFAR-10 classifiers generalize to CIFAR-10?

Summary of paper: Understand the danger of overfitting to repeatedly used datasets by creating a new test set of truly unseen images similar to CIFAR-10.

Experimental design:

- Tiny Images dataset of 80 million 32x32 images
- CIFAR-10 was created from Tiny images and consists of 60,000 images with 10 classes (equally balanced)



(a) Test Set A (b) Test Set B

Figure 1: Class-balanced random draws from the new and original test sets.[1]

- Followed "same" procedure (details in paper) to create another test dataset of 2000 images with distribution similar to CIFAR-10

Case Study 2

Do CIFAR-10 classifiers generalize to CIFAR-10?

Experimental results:

	Original Accuracy	New Accuracy	Gap	Δ Rank
shake_shake_64d_cutout [3, 4]	97.1 [96.8, 97.4]	93.0 [91.8, 94.0]	4.1	0
shake_shake_96d [4]	97.1 [96.7, 97.4]	91.9 [90.7, 93.1]	5.1	-2
shake_shake_64d [4]	97.0 [96.6, 97.3]	91.4 [90.1, 92.6]	5.6	-2
wide_resnet_28_10_cutout [3, 22]	97.0 [96.6, 97.3]	92.0 [90.7, 93.1]	5	+1
shake_drop [21]	96.9 [96.5, 97.2]	92.3 [91.0, 93.4]	4.6	+3
shake_shake_32d [4]	96.6 [96.2, 96.9]	89.8 [88.4, 91.1]	6.8	-2
darc [11]	96.6 [96.2, 96.9]	89.5 [88.1, 90.8]	7.1	-4
resnext_29_4x64d [20]	96.4 [96.0, 96.7]	89.6 [88.2, 90.9]	6.8	-2
pyramidnet_basic_110_270 [6]	96.3 [96.0, 96.7]	90.5 [89.1, 91.7]	5.9	+3
resnext_29_8x64d [20]	96.2 [95.8, 96.6]	90.0 [88.6, 91.2]	6.3	+3
wide_resnet_28_10 [22]	95.9 [95.5, 96.3]	89.7 [88.3, 91.0]	6.2	+2
pyramidnet_basic_110_84 [6]	95.7 [95.3, 96.1]	89.3 [87.8, 90.6]	6.5	0
densenet_BC_100_12 [10]	95.5 [95.1, 95.9]	87.6 [86.1, 89.0]	8	-2
neural_architecture_search [23]	95.4 [95.0, 95.8]	88.8 [87.4, 90.2]	6.6	+1
wide_resnet_tf [22]	95.0 [94.6, 95.4]	88.5 [87.0, 89.9]	6.5	+1
resnet_v2_bottleneck_164 [8]	94.2 [93.7, 94.6]	85.9 [84.3, 87.4]	8.3	-1
vgg16_keras [14, 18]	93.6 [93.1, 94.1]	85.3 [83.6, 86.8]	8.3	-1
resnet_basic_110 [7]	93.5 [93.0, 93.9]	85.2 [83.5, 86.7]	8.3	-1
resnet_v2_basic_110 [8]	93.4 [92.9, 93.9]	86.5 [84.9, 88.0]	6.9	+3
resnet_basic_56 [7]	93.3 [92.8, 93.8]	85.0 [83.3, 86.5]	8.3	0

Case Study 2

Do CIFAR-10 classifiers generalize to CIFAR-10?

Experimental results:

- significant drop in accuracy from the original test set to our new test set e.g. VGGnet and ResNet dropped from 93% to 85%
- more recent models with higher original accuracy show a smaller drop e.g. top model is still a recent Shake-Shake network with Cutout regularization (its advantage over other methods such as ResNet increased from 4% to 8%)

“In spite of adapting to the CIFAR-10 test set for several years, there has been no stagnation.” ““attacking” a test set for an extended period of time is surprisingly resilient to overfitting.”

HOWEVER “cast doubt on the robustness”

“even in benign settings, distribution shift poses a serious challenge and questions to what extent current models truly generalize”

Some questions to consider when critiquing experimental evaluation

- What is the purpose / stated goal(s) of the empirical evaluation?
- Is the experiment design reasonable given the stated goals?
- Are the stated goals achieved?
- Are appropriate baseline methods considered?
- Are appropriate evaluation metrics used?
- **Do the results account for the inherent uncertainty associated with data-driven approaches?**

High mean test accuracy doesn't make a classifier better

Accuracy of classifier

	Mean
• Classifier 1	92%
• Classifier 2	87%

High mean test accuracy doesn't make a classifier better

Accuracy of classifier

	Mean	Std
• Classifier 1	92%	15%
• Classifier 2	87%	5%

High mean test accuracy doesn't make a classifier better

Accuracy of classifier

	Mean	Std	Range
• Classifier 1	92%	15%	77-100
• Classifier 2	87%	5%	82-92

Case Study 1

ImageNet Classification with Deep Convolutional Neural Networks

Experimental results:

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**

Table 1: Comparison of results on ILSVRC-2010 test set. In *italics* are best results achieved by others.

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

Table 2: Comparison of error rates on ILSVRC-2012 validation and test sets. In *italics* are best results achieved by others. Models with an asterisk* were "pre-trained" to classify the entire ImageNet 2011 Fall release. See Section 6 for details.

Case Study 2

Do CIFAR-10 classifiers generalize to CIFAR-10?

Experimental results:

	Original Accuracy	New Accuracy	Gap	Δ Rank
shake_shake_64d_cutout [3, 4]	97.1 [96.8, 97.4]	93.0 [91.8, 94.0]	4.1	0
shake_shake_96d [4]	97.1 [96.7, 97.4]	91.9 [90.7, 93.1]	5.1	-2
shake_shake_64d [4]	97.0 [96.6, 97.3]	91.4 [90.1, 92.6]	5.6	-2
wide_resnet_28_10_cutout [3, 22]	97.0 [96.6, 97.3]	92.0 [90.7, 93.1]	5	+1
shake_drop [21]	96.9 [96.5, 97.2]	92.3 [91.0, 93.4]	4.6	+3
shake_shake_32d [4]	96.6 [96.2, 96.9]	89.8 [88.4, 91.1]	6.8	-2
darc [11]	96.6 [96.2, 96.9]	89.5 [88.1, 90.8]	7.1	-4
resnext_29_4x64d [20]	96.4 [96.0, 96.7]	89.6 [88.2, 90.9]	6.8	-2
pyramidnet_basic_110_270 [6]	96.3 [96.0, 96.7]	90.5 [89.1, 91.7]	5.9	+3
resnext_29_8x64d [20]	96.2 [95.8, 96.6]	90.0 [88.6, 91.2]	6.3	+3
wide_resnet_28_10 [22]	95.9 [95.5, 96.3]	89.7 [88.3, 91.0]	6.2	+2
pyramidnet_basic_110_84 [6]	95.7 [95.3, 96.1]	89.3 [87.8, 90.6]	6.5	0
densenet_BC_100_12 [10]	95.5 [95.1, 95.9]	87.6 [86.1, 89.0]	8	-2
neural_architecture_search [23]	95.4 [95.0, 95.8]	88.8 [87.4, 90.2]	6.6	+1
wide_resnet_tf [22]	95.0 [94.6, 95.4]	88.5 [87.0, 89.9]	6.5	+1
resnet_v2_bottleneck_164 [8]	94.2 [93.7, 94.6]	85.9 [84.3, 87.4]	8.3	-1
vgg16_keras [14, 18]	93.6 [93.1, 94.1]	85.3 [83.6, 86.8]	8.3	-1
resnet_basic_110 [7]	93.5 [93.0, 93.9]	85.2 [83.5, 86.7]	8.3	-1
resnet_v2_basic_110 [8]	93.4 [92.9, 93.9]	86.5 [84.9, 88.0]	6.9	+3
resnet_basic_56 [7]	93.3 [92.8, 93.8]	85.0 [83.3, 86.5]	8.3	0

Purpose often dictates validity of classifier

Accuracy of classifier

	Mean	Std	Range
• Classifier 1	92%	15%	77-100
• Classifier 2	87%	5%	82-92

Which classifier would you choose when recommending movies?

Which classifier would you choose when diagnosing serious illness?

Purpose often dictates validity of regressor

Accuracy of regressor

	MSE
• Regressor 1	25
• Regressor 2	0.0001

Purpose often dictates validity of regressor

Accuracy of regressor

	MSE	Task
• Regressor 1	25	Predict age of a person
• Regressor 2	0.0001	Predict proportion of lead in water

MSE vs. MAE

Units important

Some questions to consider when critiquing experimental evaluation

- What is the purpose / stated goal(s) of the empirical evaluation?
- Is the experiment design reasonable given the stated goals?
- Are the stated goals achieved?
- Are appropriate baseline methods considered?
- Are appropriate evaluation metrics used?
- Do the results account for the inherent uncertainty associated with data-driven approaches?
- **Are the written discussions and conclusions corroborated by the actual empirical results?**

Interpreting ‘correct’ results correctly is important too

- Confounding variables

  Given data from a surveillance camera, an ML algorithm could predict with high accuracy when a subway is busy. Hence, it has learnt to detect crowd.

  Given images of US and Russian tanks, an ML algorithm could classify them with high accuracy. Hence, it learnt to distinguish between their salient capabilities.

Interpreting ‘correct’ results correctly is important too

Automated Inference on Criminality using Face Images

ML algorithms can classify criminals based on face images. “… find some discriminating structural features for predicting criminality, such as lip curvature, eye inner corner distance, and the so-called nose-mouth angle ….”

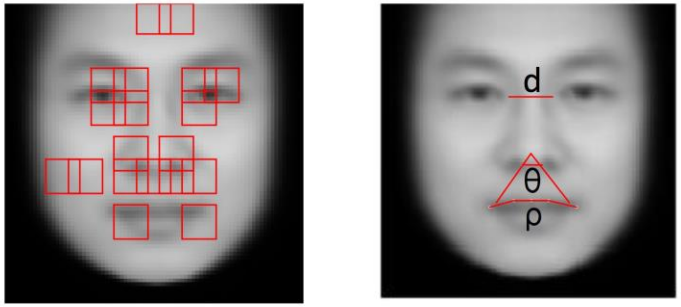


Figure 4. (a) FGM results; (b) Three discriminative features ρ, d and θ.

Deep Neural Networks Are More Accurate Than Humans at Detecting Sexual Orientation From Facial Images

“We show that faces contain much more information about sexual orientation than can be perceived and interpreted by the human brain.”

Interpreting ‘correct’ results correctly is important too

- Correlation vs Causation

London taxi drivers: A survey has pointed out a positive and significant correlation between the number of accidents and wearing coats. They concluded that coats could hinder movements of drivers and be the cause of accidents. A new law was prepared to prohibit drivers from wearing coats when driving.

Finally another study pointed out that people wear coats when it rains…

Some questions to consider when critiquing experimental evaluation

- What is the purpose / stated goal(s) of the empirical evaluation?
- Is the experiment design reasonable given the stated goals?
- Are the stated goals achieved?
- Are appropriate baseline methods considered?
- Are appropriate evaluation metrics used?
- Do the results account for the inherent uncertainty associated with data-driven approaches?
- Are the written discussions and conclusions corroborated by the actual empirical results?
- **Are the empirical results reproducible?**

Can experiments be reproduced?

- All model choices mentioned?
  - Model family, Step-size, batch-size, initialization, order of cross-validation, training/validation/test/hold-out set size, …
- Experimental platform details?
  - Which GPUs, CPUs, memory, …
- Data and code availability?
  - Random seeds?
  - Is code itself deterministic given random seeds?

Case Study 1

ImageNet Classification with Deep Convolutional Neural Networks

+ Details of implementation

+ Dataset public

+ Implementation public

http://code.google.com/p/cuda-convnet/