Evaluation and Interpretability in Deep Neural Networks

AMIA 2018

John Lalor[1], Abhyuday Jagannatha[1], Hong Yu[2]
[1]CICS, UMass Amherst
[2]Department of Computer Science, UMass Lowell

November 3, 2018

Disclosure

We have no relevant relationships with commercial interests to disclose

Learning Objectives

After participating in this session the learner should be better able to:

- Apply traditional evaluation metrics to deep learning models
- Design new test sets for deep learning models
- Understand specific types of interpretability and identify when they are required

About me

John Lalor

PhD Candidate at UMass Amherst CICS

My research: Deep learning evaluation, health informatics, leveraging uncertainty in model training

---

http://people.cs.umass.edu/~lalor

About me

John Lalor

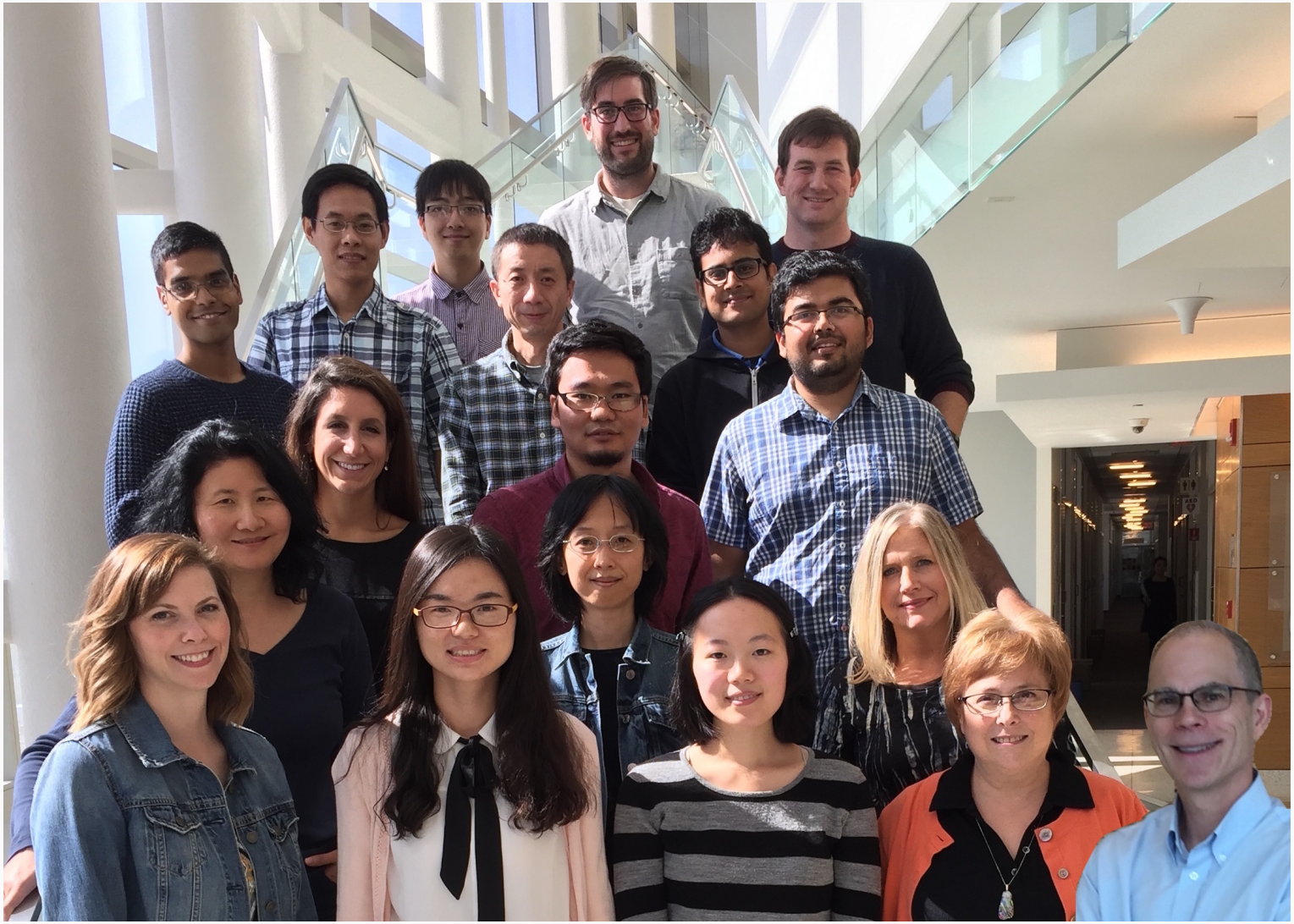
PhD Candidate at UMass Amherst CICS

My research: Deep learning evaluation, health informatics, leveraging uncertainty in model training

I'm on the job market!

`http://people.cs.umass.edu/~lalor`

Bio-NLP Lab



http://bio-nlp.org

Join us!

Post-doc position available in the lab

Email: hong.yu@umassmed.edu

Talk outline

Traditional model evaluation

Hands-on: Evaluation

New evaluation methods

Interpretability

Hands-on: Interpretability

Today

Evaluation and Interpretability in Deep Neural Networks (DNNs)

Today

Evaluation and Interpretability in Deep Neural Networks (DNNs)

How to evaluate DNN performance

Today

Evaluation and Interpretability in Deep Neural Networks (DNNs)

How to interpret outputs in order to understand model learning

Table of contents

Traditional model evaluation

Hands-on: Evaluation

New evaluation methods

Interpretability

Hands-on: Interpretability

Supervised learning

Data: (x, y)

Supervised learning

Data: (x, y)

Learn a function $h(x)$ that maps x (inputs) to y (outputs).

Supervised learning

Data: (x, y)

Learn a function $h(x)$ that maps x (inputs) to y (outputs).

- Output can be:

Supervised learning

Data: (x, y)

Learn a function $h(x)$ that maps x (inputs) to y (outputs).

- Output can be:
  - Real-valued $h(x) \in \mathbb{R}$

Supervised learning

Data: (x, y)

Learn a function $h(x)$ that maps x (inputs) to y (outputs).

- Output can be:
  - Real-valued $h(x) \in \mathbb{R}$
  - 1 of C classes: $h(x) \in 1, 2, \ldots, C$

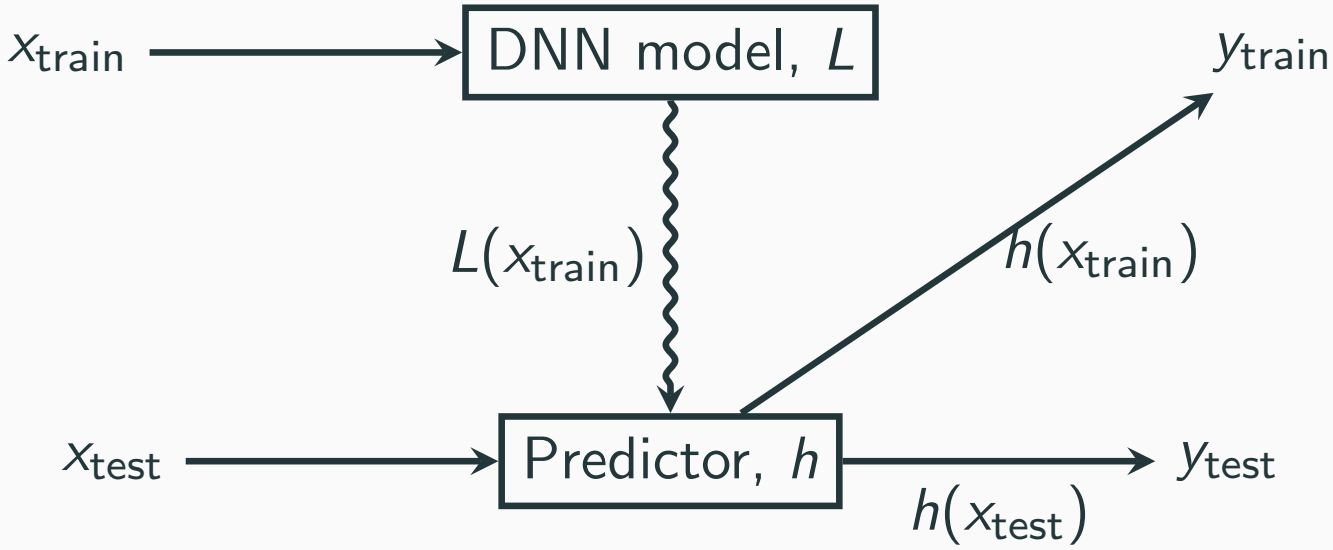
Supervised learning

Data: (x, y)

Learn a function $h(x)$ that maps x (inputs) to y (outputs).

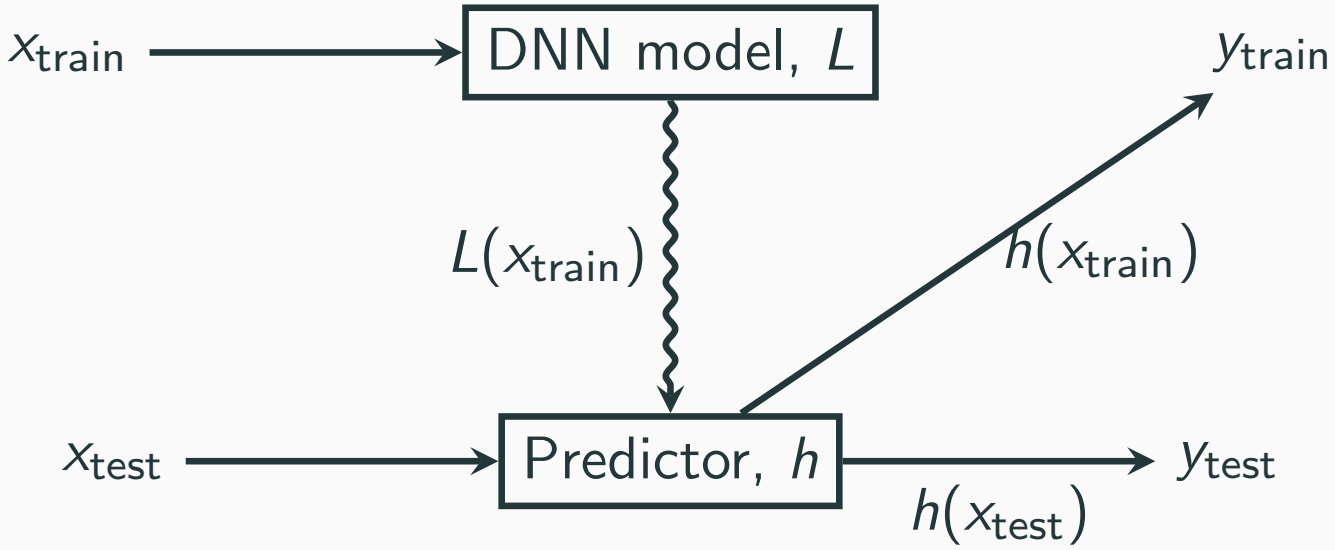
- Output can be:
  - Real-valued $h(x) \in \mathbb{R}$
  - 1 of C classes: $h(x) \in 1, 2, \ldots, C$
  - Probability estimate: $h(x) \in [0, 1]$

Accuracy



Petar Velikovi, https://github.com/PetarV-/TikZ, 2018

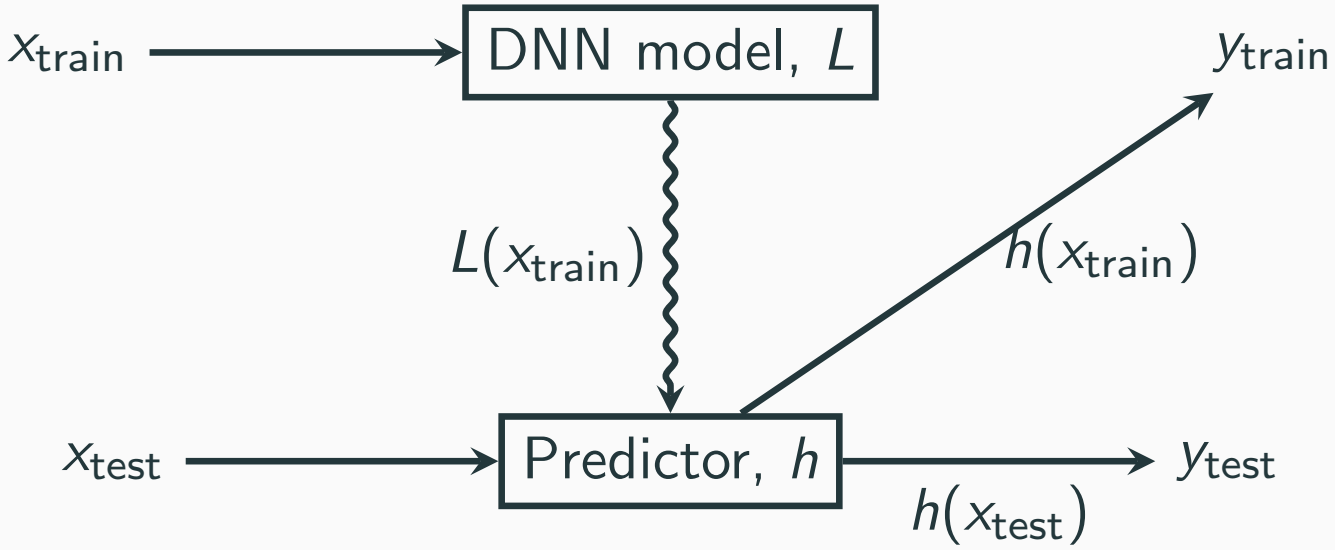
Accuracy



Training error: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[h(x^i_{\text{train}}) \neq y^i_{\text{train}}]$

Petar Velikovi, https://github.com/PetarV-/TikZ, 2018

Accuracy



Training error: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[h(x^i_{\text{train}}) \neq y^i_{\text{train}}]$

Test error: $\frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[h(x^i_{\text{test}}) \neq y^i_{\text{test}}]$

Petar Velikovi, https://github.com/PetarV-/TikZ, 2018

Precision, recall, and F1

Important for when you have imbalanced data (disease classification, anomaly detection, etc.)

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Precision, recall, and F1

Important for when you have imbalanced data (disease classification, anomaly detection, etc.)

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Precision: $\frac{\text{TP}}{\text{TP+FP}}$

Precision, recall, and F1

Important for when you have imbalanced data (disease classification, anomaly detection, etc.)

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Precision: $\frac{\text{TP}}{\text{TP+FP}}$

Recall: $\frac{\text{TP}}{\text{TP+FN}}$

Precision, recall, and F1

Important for when you have imbalanced data (disease classification, anomaly detection, etc.)

		Predicted	
		Negative	Positive
Actual	Negative	True Negative	False Positive
	Positive	False Negative	True Positive

Precision: $\frac{\text{TP}}{\text{TP}+\text{FP}}$

Recall: $\frac{\text{TP}}{\text{TP}+\text{FN}}$

F1-score: $2 \times \frac{P \times R}{P+R}$

Example

Positive: 100

Negative: 900

		Predicted	
		Negative	Positive
Actual	Negative	800	100
	Positive	20	80

Example

Positive: 100

Negative: 900

		Predicted	
		Negative	Positive
Actual	Negative	800	100
	Positive	20	80

Accuracy: $\frac{800+80}{1000} = 0.88$

Example

Positive: 100

Negative: 900

		Predicted	
		Negative	Positive
Actual	Negative	800	100
	Positive	20	80

Accuracy: $\frac{800+80}{1000} = 0.88$

Precision: $\frac{80}{80+100} = 0.444$

Example

Positive: 100

Negative: 900

		Predicted	
		Negative	Positive
Actual	Negative	800	100
	Positive	20	80

Accuracy: $\frac{800+80}{1000} = 0.88$

Precision: $\frac{80}{80+100} = 0.444$

Recall: $\frac{80}{80+20} = 0.80$

Example

Positive: 100

Negative: 900

		Predicted	
		Negative	Positive
Actual	Negative	800	100
	Positive	20	80

Accuracy: $\frac{800+80}{1000} = 0.88$

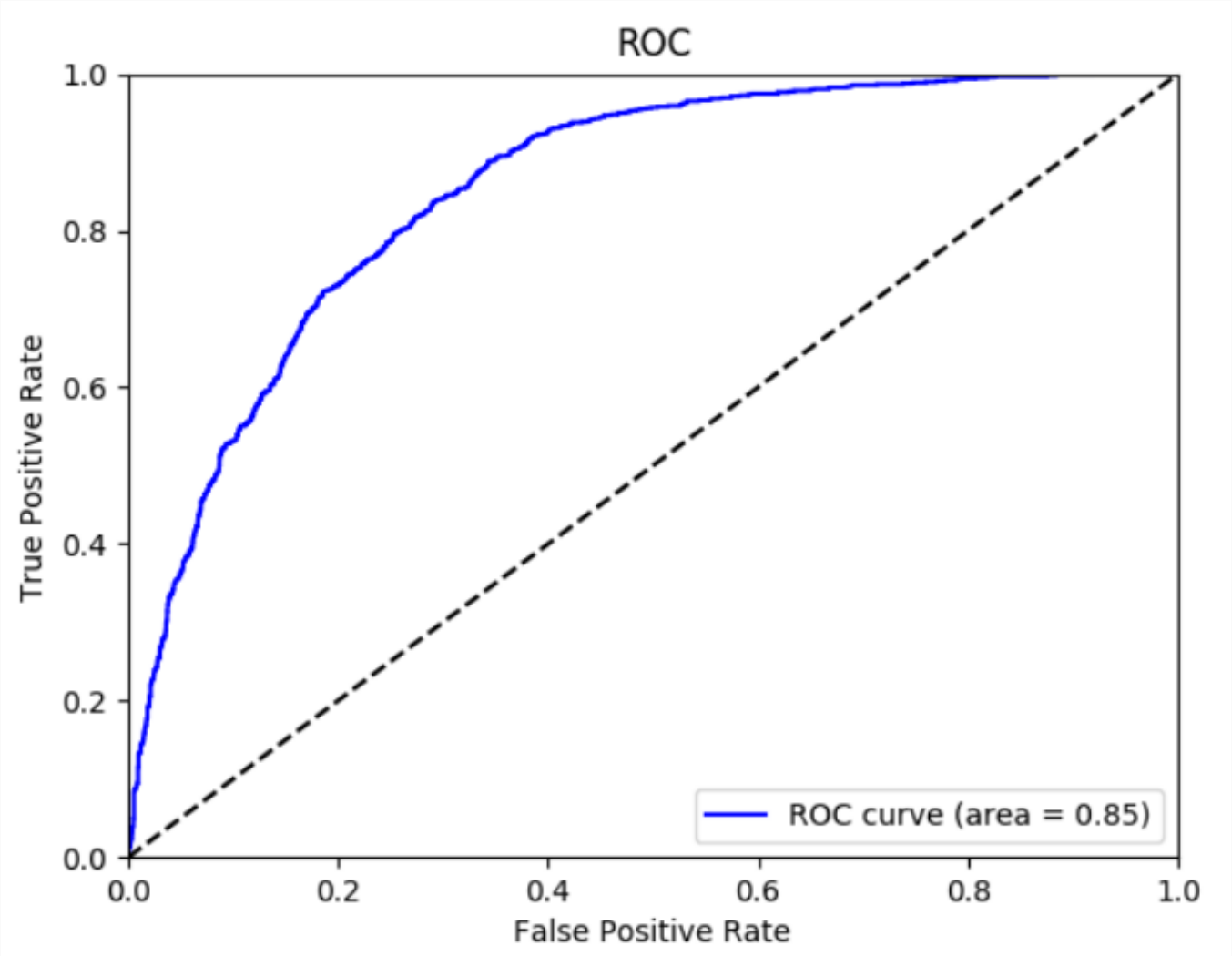
Precision: $\frac{80}{80+100} = 0.444$

Recall: $\frac{80}{80+20} = 0.80$

F1-score: $2 \times \frac{P \times R}{P+R} = 0.571$

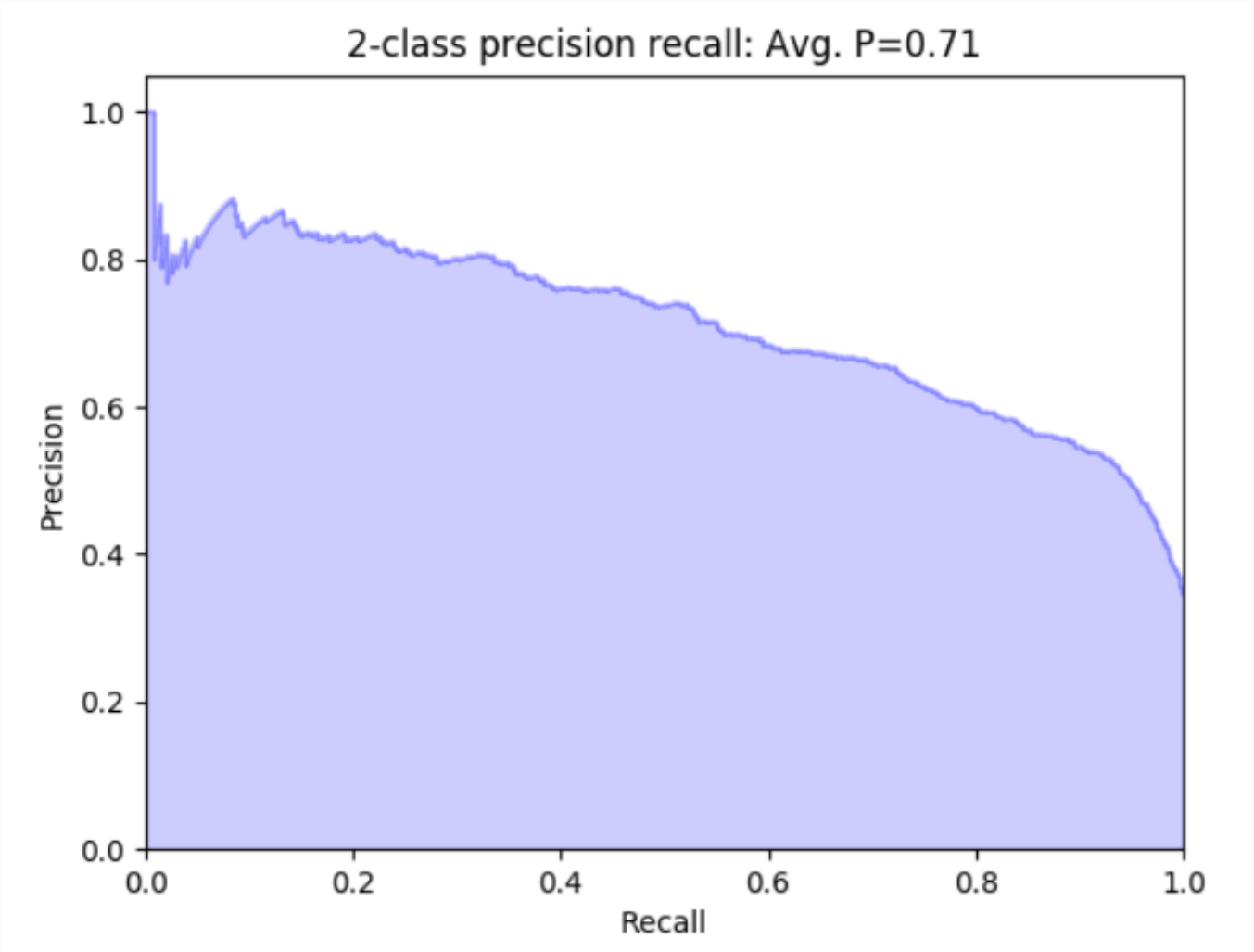
Receiver operator characteristic (ROC)

Measures performance as the threshold is varied

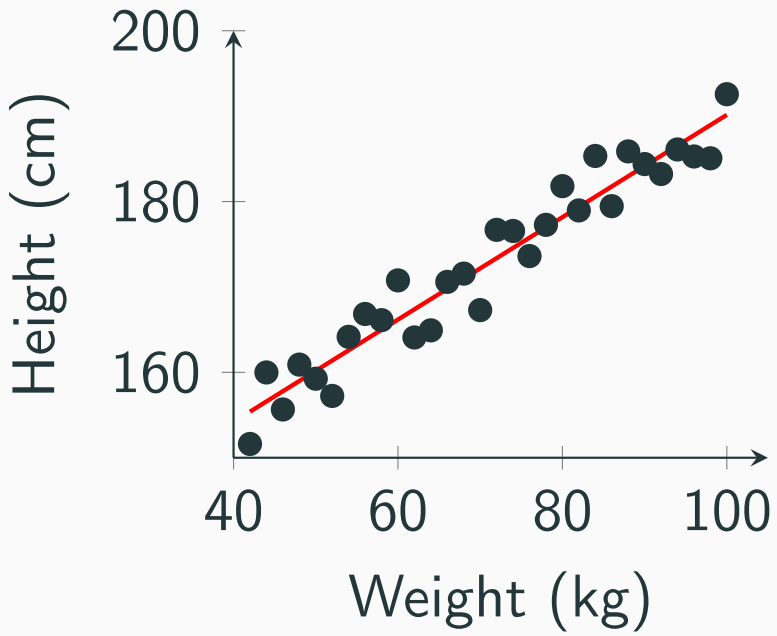


Precision-recall curve

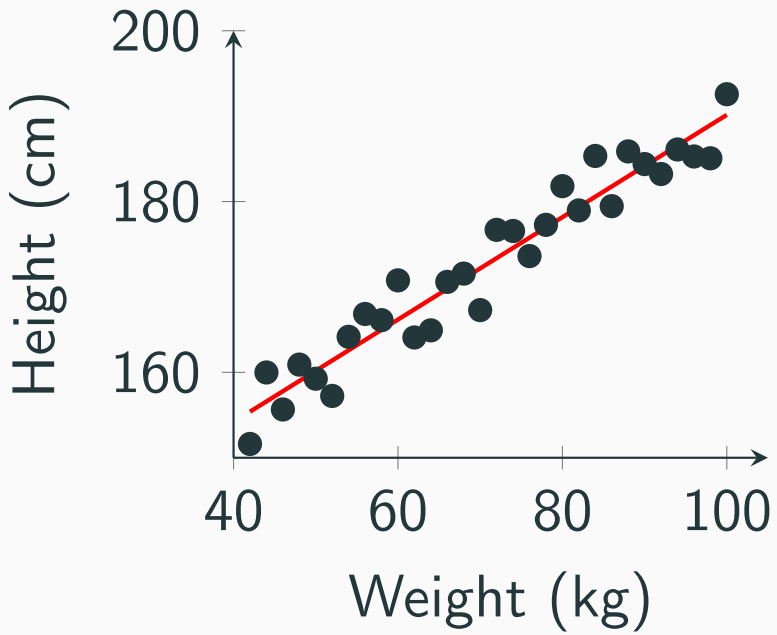
Tradeoff between precision and recall



Regression metrics



Regression metrics



Training error:

$$\frac{1}{N}\sum_{n=1}^{N}(h(x^n_{\text{train}}) - y^n_{\text{train}})^2$$

Testing error:

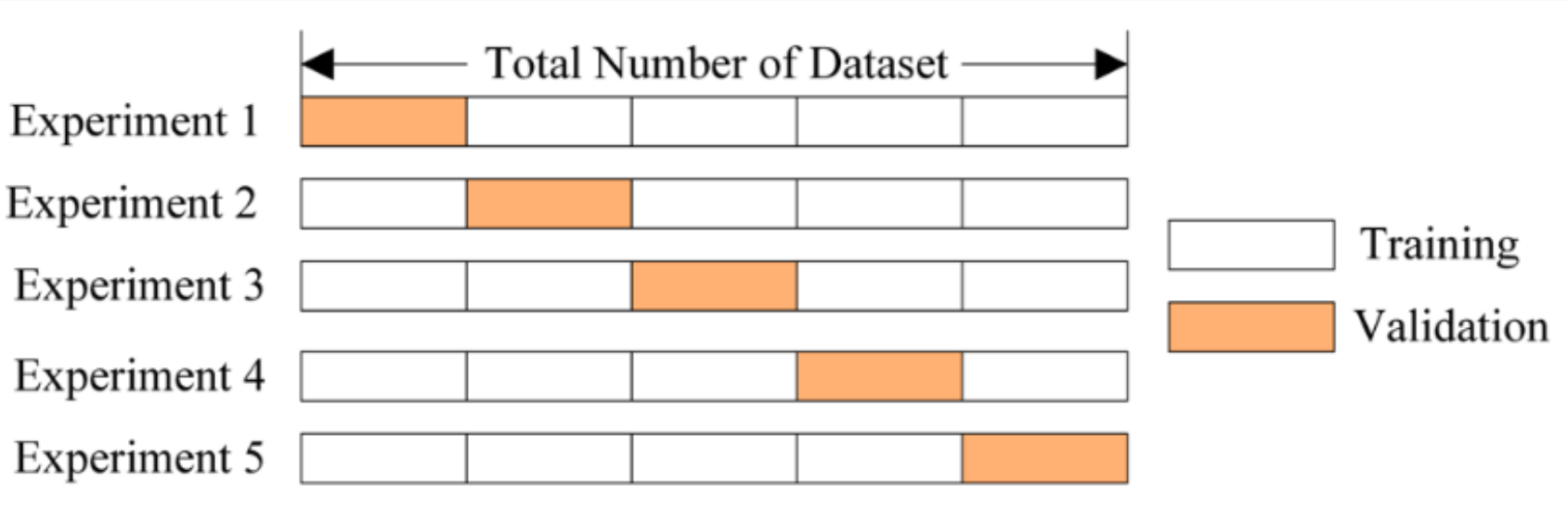
$$\frac{1}{N}\sum_{n=1}^{N}(h(x^n_{\text{test}}) - y^n_{\text{test}})^2$$

Cross-validation

Split data into k "folds:" $x_1, x_2, \ldots, x_k$

Train model on $k - 1$ folds, test on the last fold

Data will be used for training $k - 1$ times and once for testing



https://www.kaggle.com/dansbecker/cross-validation

Deep learning considerations

Can train on huge data sets (pros & cons)

Can require a lot of computing power (e.g. GPUs)

Model comparison

Use a previously released dataset for training/testing

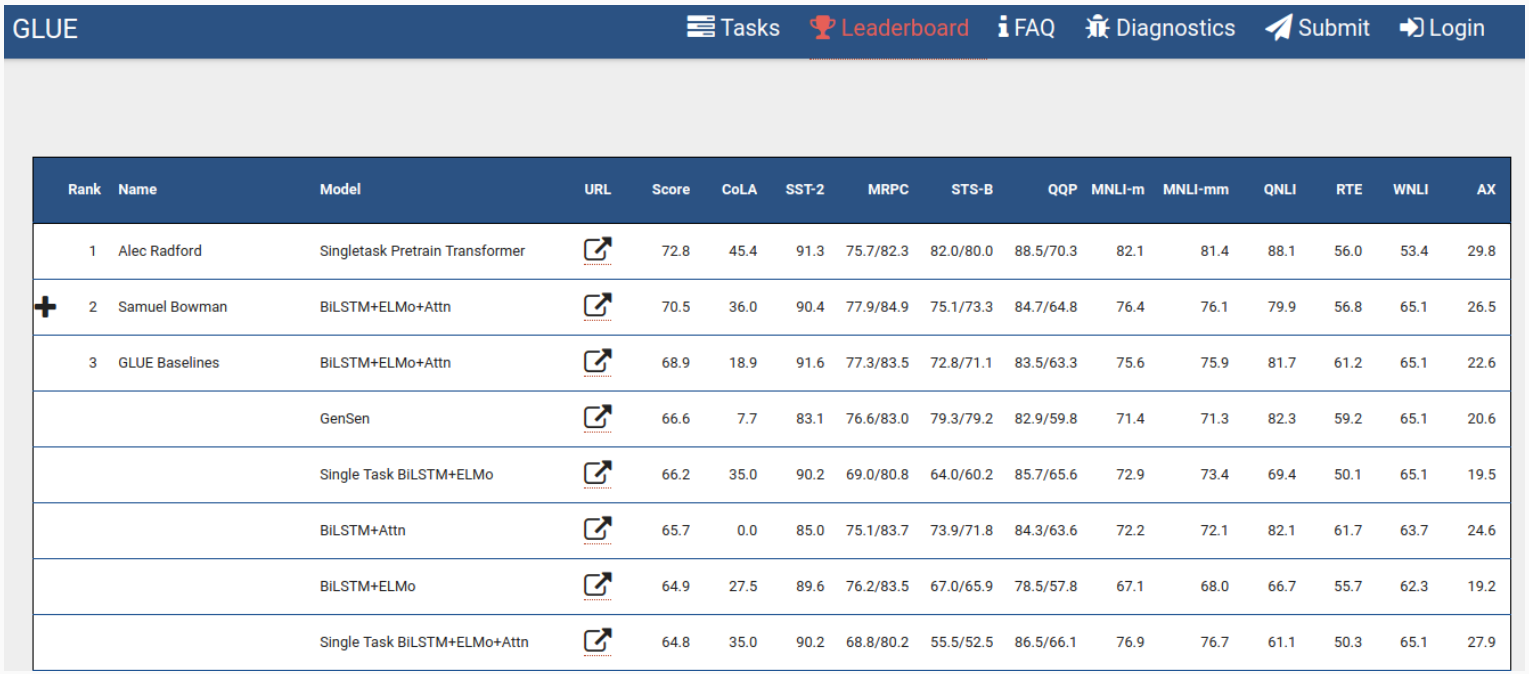
Compare test results to other models (leaderboard)

Model comparison

Use a previously released dataset for training/testing

Compare test results to other models (leaderboard)

General Language Understanding Evaluation: sentiment, paraphrase, NLI



GLUE

Tasks | Leaderboard | FAQ | Diagnostics | Submit | Login

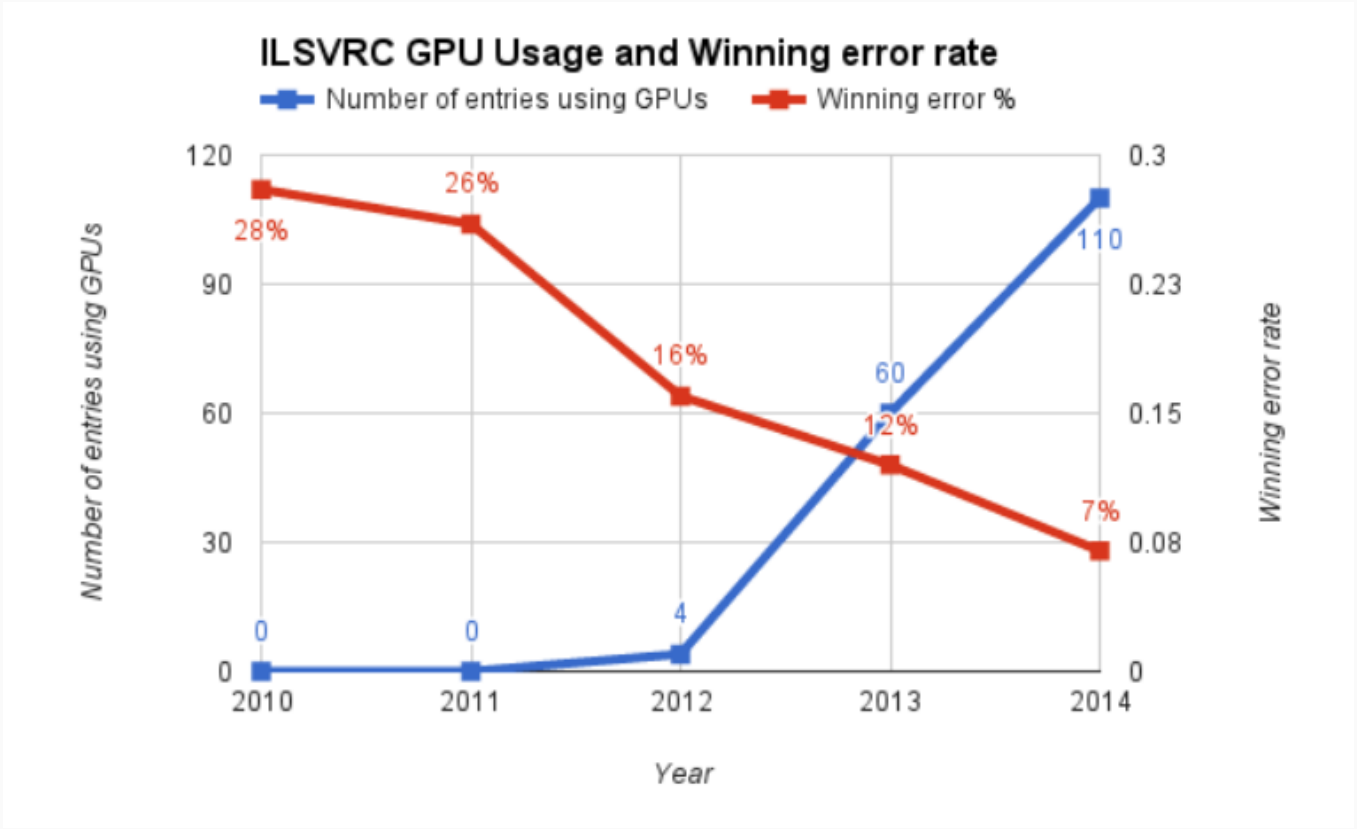
Rank	Name	Model	URL	Score	CoLA	SST-2	MRPC	STS-B	QQP	MNLI-m	MNLI-mm	QNLI	RTE	WNLI	AX
1	Alec Radford	Singletask Pretrain Transformer		72.8	45.4	91.3	75.7/82.3	82.0/80.0	88.5/70.3	82.1	81.4	88.1	56.0	53.4	29.8
2	Samuel Bowman	BiLSTM+ELMo+Attn		70.5	36.0	90.4	77.9/84.9	75.1/73.3	84.7/64.8	76.4	76.1	79.9	56.8	65.1	26.5
3	GLUE Baselines	BiLSTM+ELMo+Attn		68.9	18.9	91.6	77.3/83.5	72.8/71.1	83.5/63.3	75.6	75.9	81.7	61.2	65.1	22.6
		GenSen		66.6	7.7	83.1	76.6/83.0	79.3/79.2	82.9/59.8	71.4	71.3	82.3	59.2	65.1	20.6
		Single Task BiLSTM+ELMo		66.2	35.0	90.2	69.0/80.8	64.0/60.2	85.7/65.6	72.9	73.4	69.4	50.1	65.1	19.5
		BiLSTM+Attn		65.7	0.0	85.0	75.1/83.7	73.9/71.8	84.3/63.6	72.2	72.1	82.1	61.7	63.7	24.6
		BiLSTM+ELMo		64.9	27.5	89.6	76.2/83.5	67.0/65.9	78.5/57.8	67.1	68.0	66.7	55.7	62.3	19.2
		Single Task BiLSTM+ELMo+Attn		64.8	35.0	90.2	68.8/80.2	55.5/52.5	86.5/66.1	76.9	76.7	61.1	50.3	65.1	27.9

https://gluebenchmark.com/

Model comparison

Use a previously released dataset for training/testing

Compare test results to other models (leaderboard)



https://beamandrew.github.io/

Reliability

Training a DNN involves an element of randomness

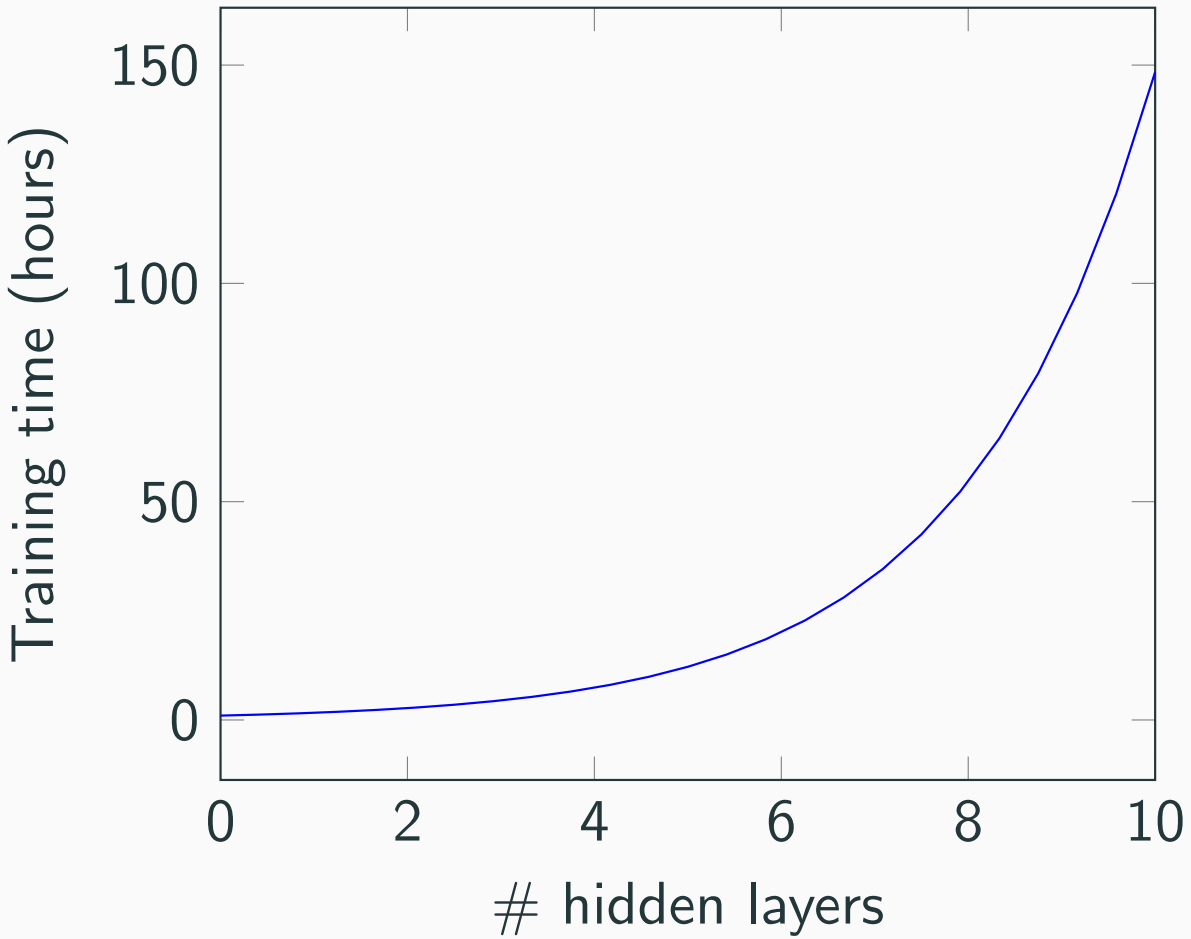
Does training model on multiple runs lead to similar results?

If you can run the model multiple times, do so!

Other considerations

Training time/Hardware

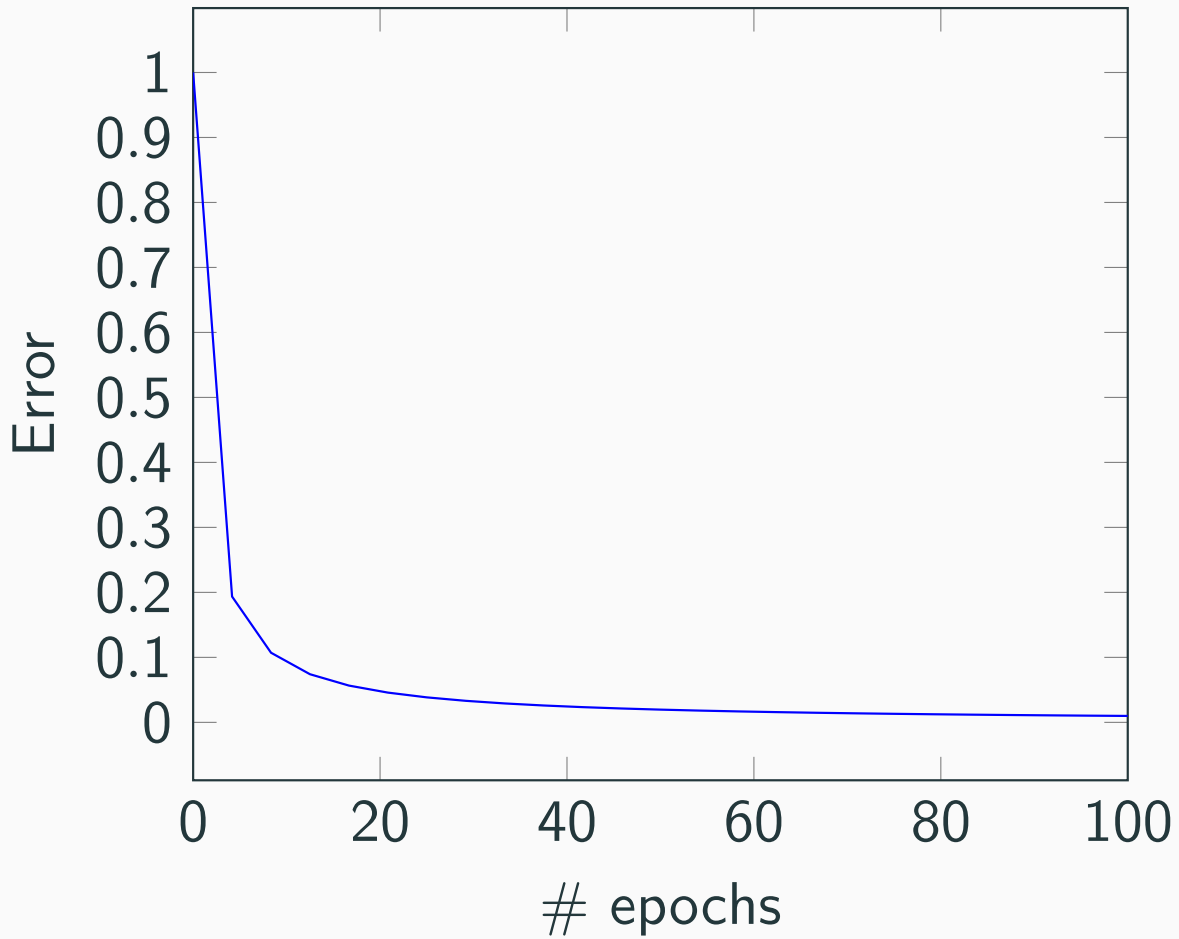
Having the best model isn't useful if it takes too long to train



Other considerations

Learning rate

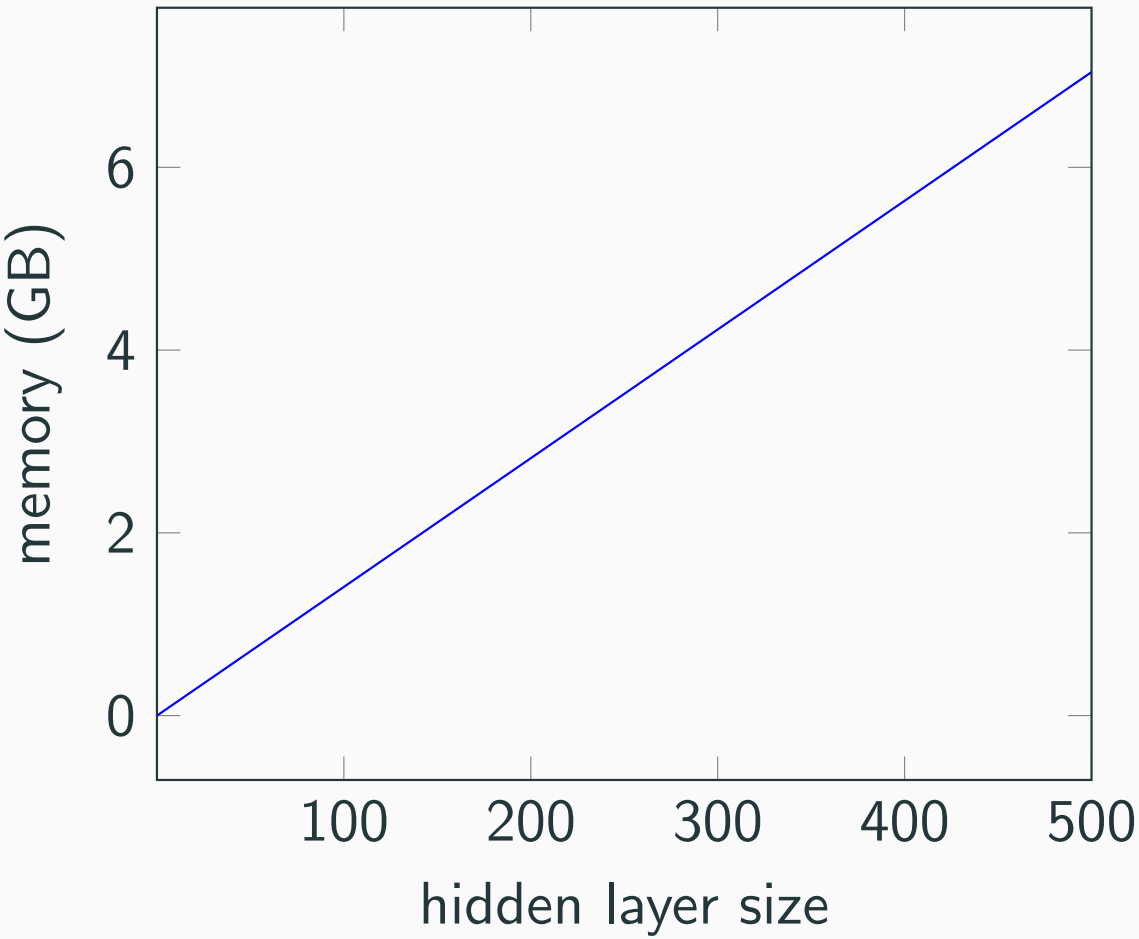
How quickly can the model generalize (# epochs)?



Other considerations

Portability

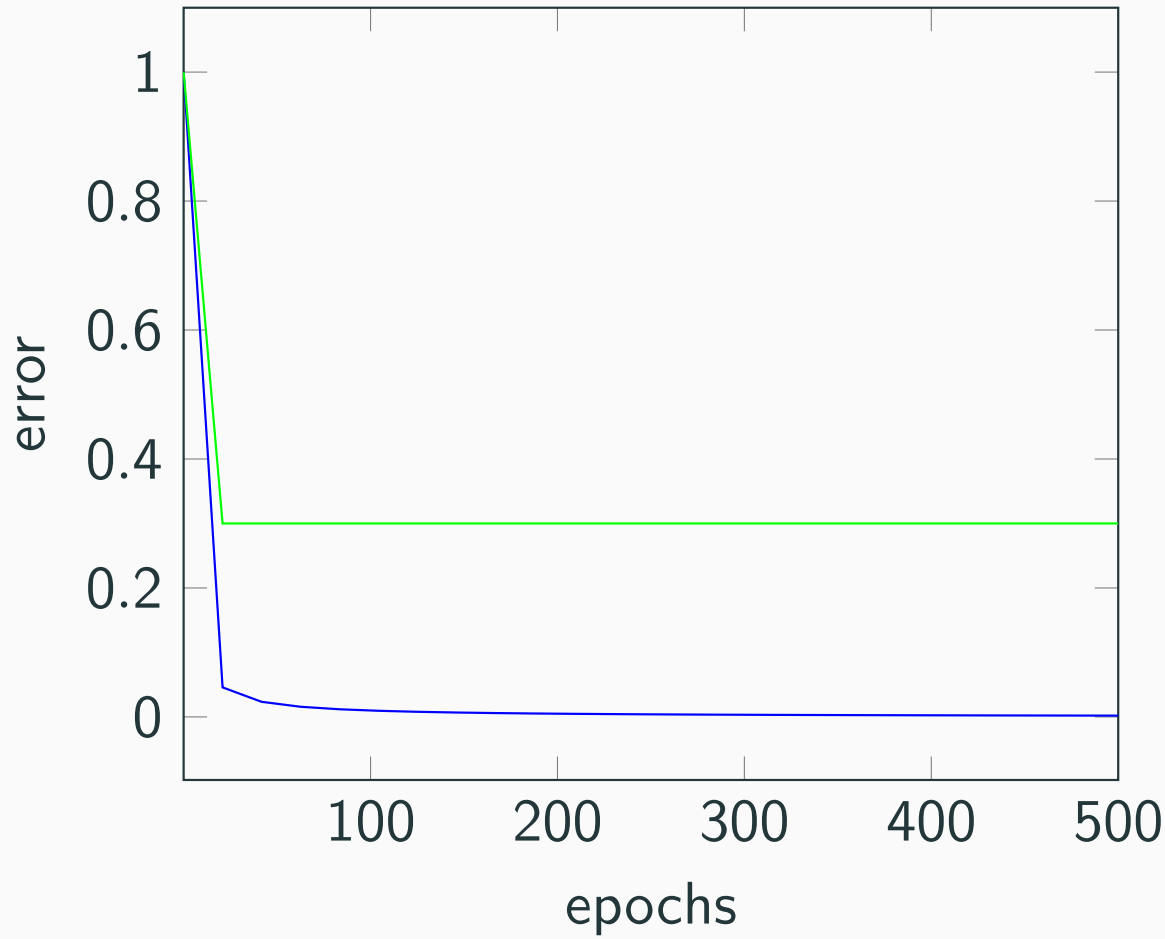
Can the model be deployed, or is it too large?



Other considerations

Overfitting

High-capacity models have risk of memorizing training data.



Unsupervised Evaluation metrics

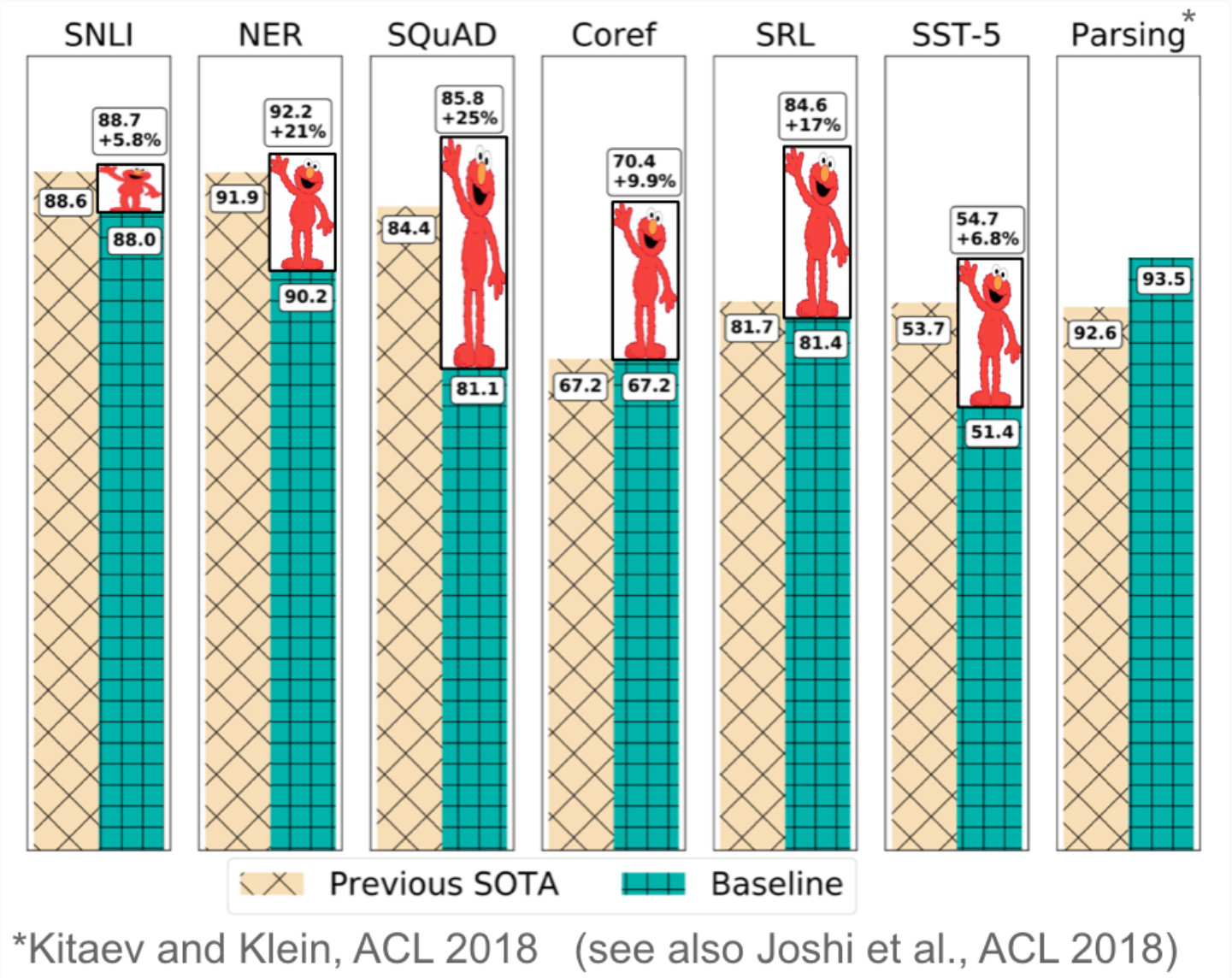
Probability metrics: log-likelihood, perplexity, etc.

Language modeling example

Input to downstream supervised learning tasks

What are you trying to learn?

Unsupervised learning



Source: Matthew Peters

How does it look?

DNNs can generate output, not just classify examples.

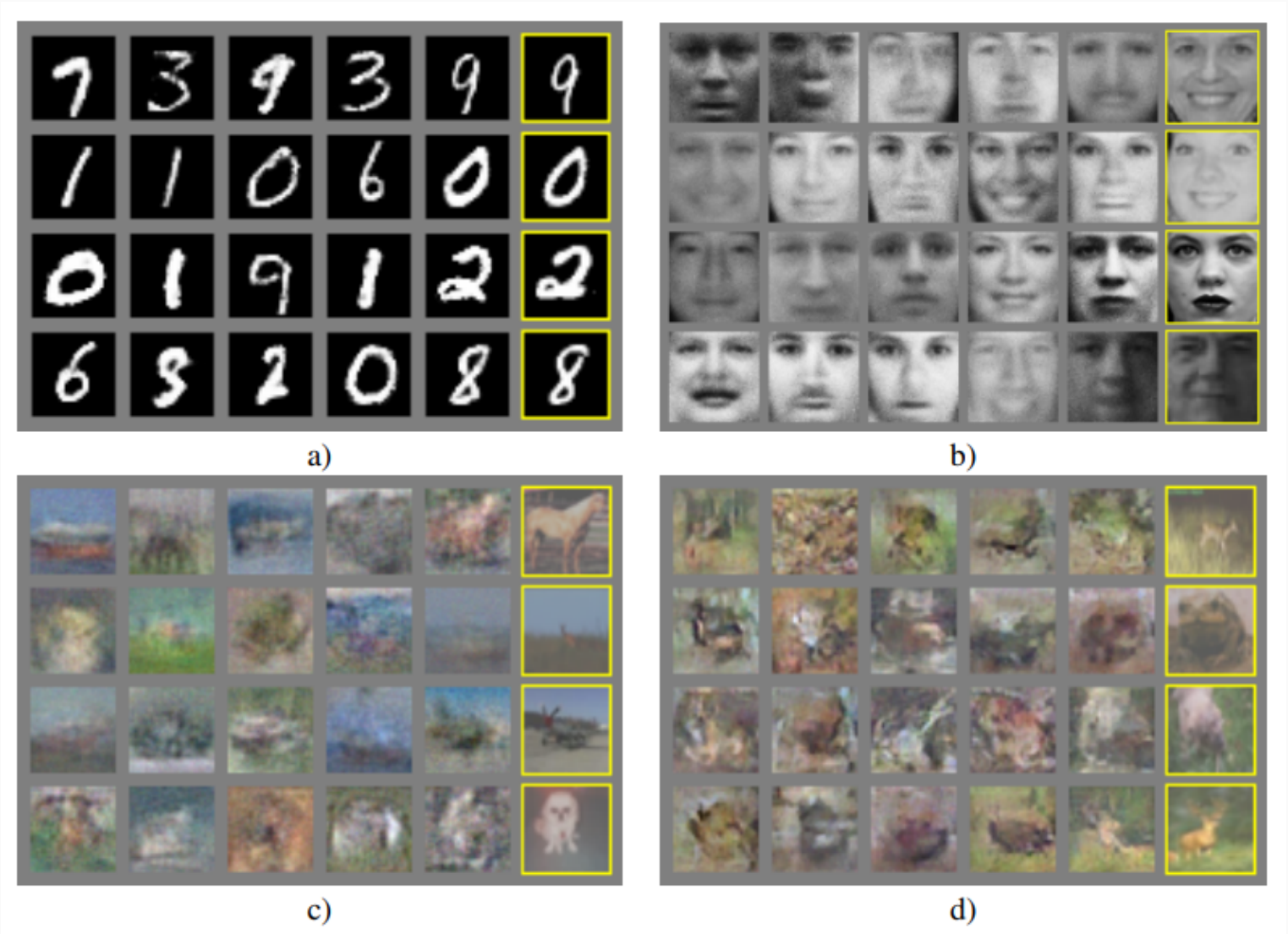
Generated outputs are harder to evaluate.

Images: visual inspection

Language: Grammatical correctness

Often requires human judgement for quality

GAN outputs



Goodfellow, Ian, et al. Generative adversarial nets. NIPS 2014.

Text generation

Table 2. Sentences generated by textGAN.

a	we show the joint likelihood estimator (in a large number of estimating variables embedded on the subspace learning) .
b	this problem achieves less interesting choices of convergence guarantees on turing machine learning .
c	in hidden markov relational spaces , the random walk feature decomposition is unique generalized parametric mappings.
d	i see those primitives specifying a deterministic probabilistic machine learning algorithm .
e	i wanted in alone in a gene expression dataset which do n't form phantom action values .
f	as opposite to a set of fuzzy modelling algorithm , pruning is performed using a template representing network structures .

Zhang, Yizhe, et al. Adversarial Feature Matching for Text Generation. ICML 2017.

Table of contents

Traditional model evaluation

Hands-on: Evaluation

New evaluation methods

Interpretability

Hands-on: Interpretability

Getting the activity code

http://jplalor.github.io/amia18

Table of contents

Traditional model evaluation

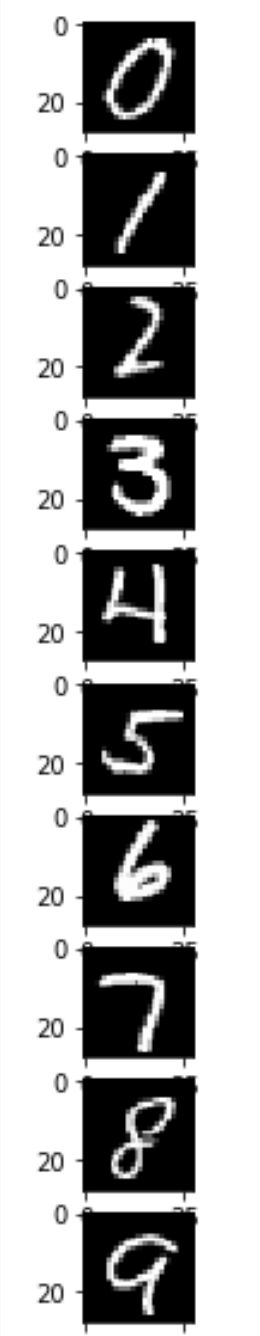
Hands-on: Evaluation

New evaluation methods

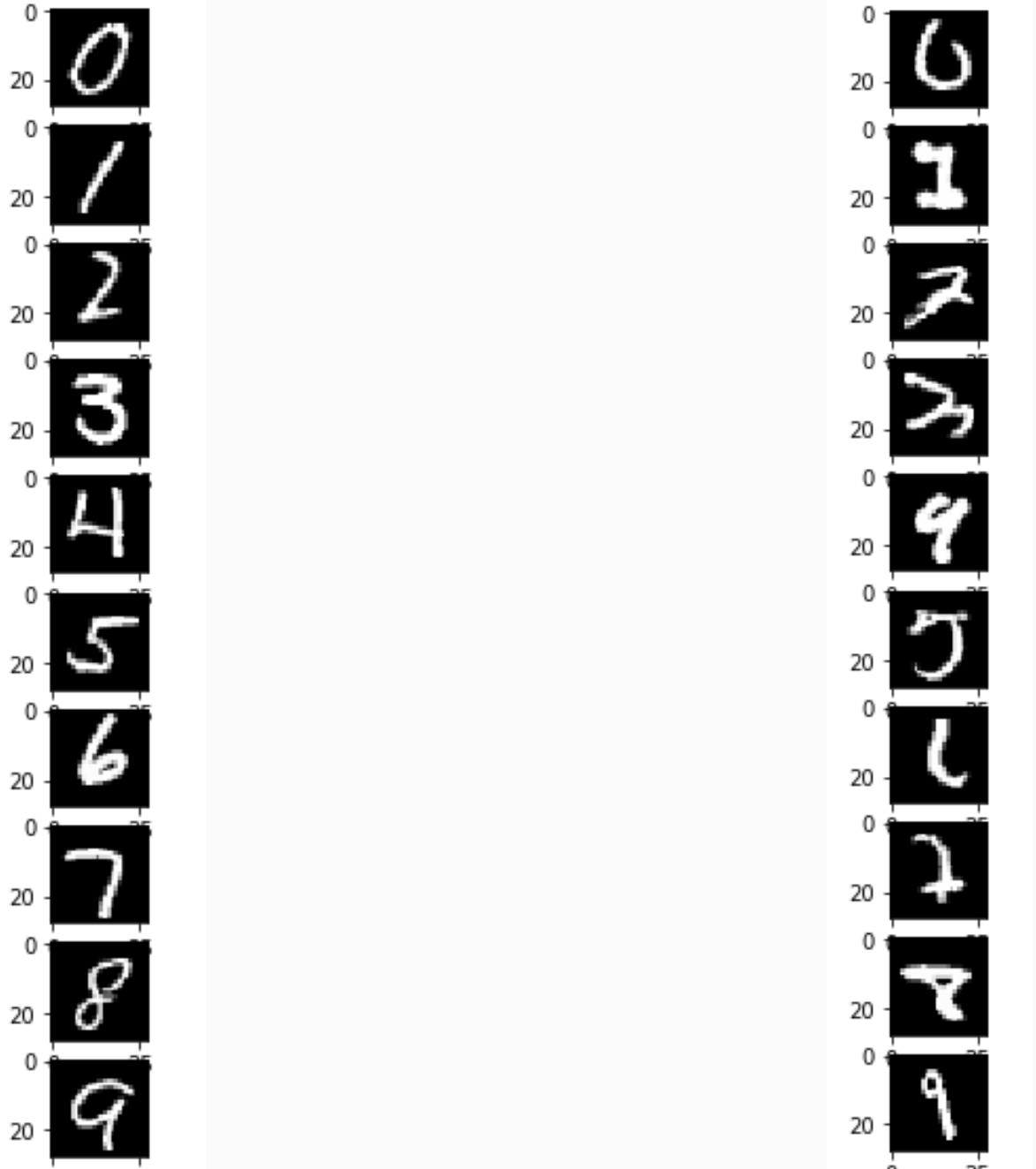
Interpretability

Hands-on: Interpretability

Motivation: digit classification



Motivation: digit classification



New methods

Ongoing research question

Methods such as Item Response Theory to compare models based on outputs to specific test set items

Adversarial examples to identify blind spots in the model

Examples

Natural language inference (NLI)

Premise	Hypothesis	Label	Difficulty
A little girl eating a sucker	A child eating candy	Entailment	-2.74
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction	0.51
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral	-1.92
Nine men wearing tuxedos sing	Nine women wearing dresses sing	Contradiction	0.08

Sentiment Analysis (SA)

Phrase	Label	Difficulty
The stupidest, most insulting movie of 2002's first quarter.	Negative	-2.46
Still, it gets the job done - a sleepy afternoon rental.	Negative	1.78
An endlessly fascinating, landmark movie that is as bold as anything the cinema has seen in years.	Positive	-2.27
Perhaps no picture ever made has more literally showed that the road to hell is paved with good intentions.	Positive	2.05

John P Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. EMNLP 2018.

Motivation

What can we do?

Motivation

What can we do?

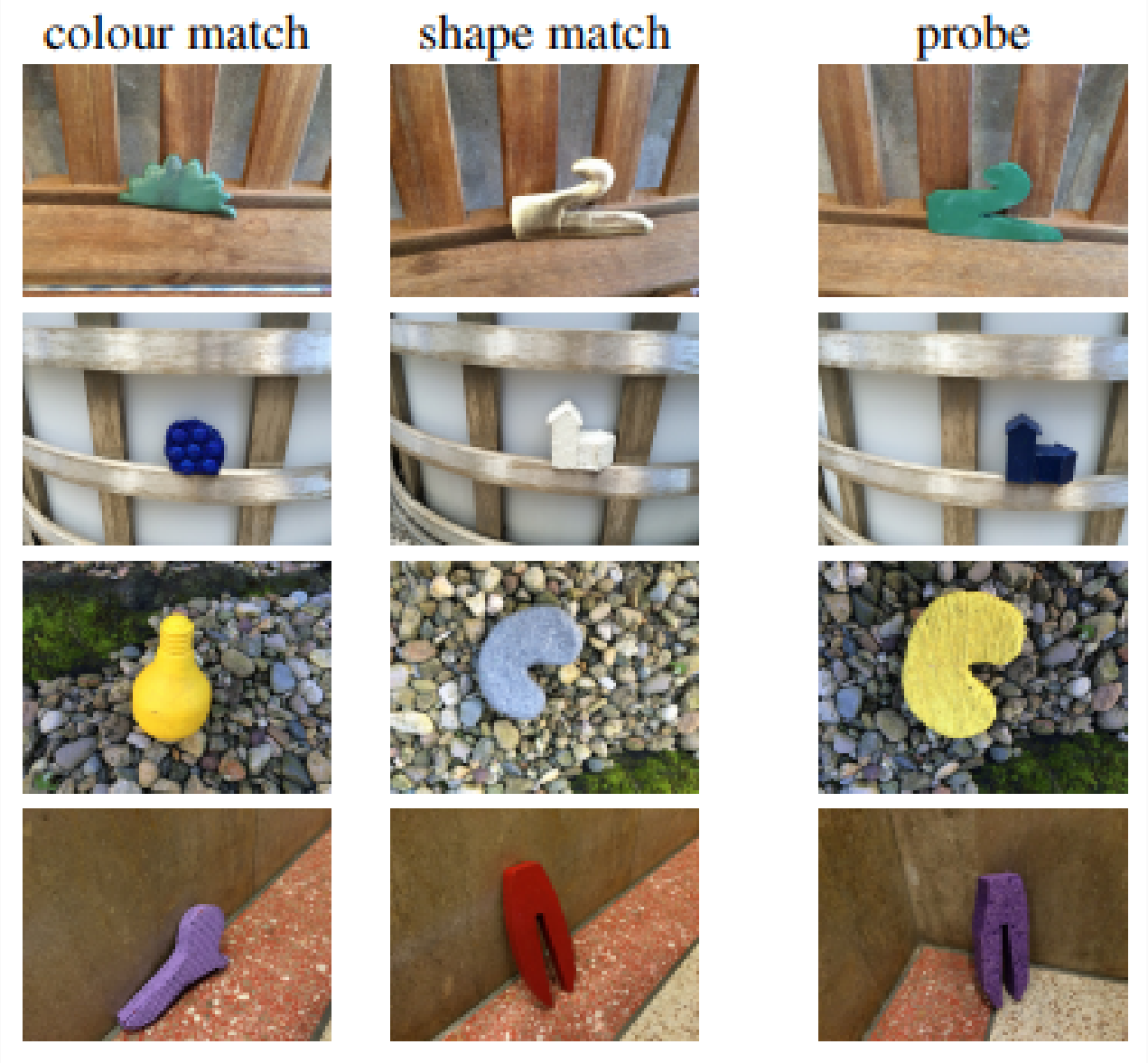
- Better training: more robust models

Motivation

What can we do?

- Better training: more robust models
- Better evaluation: see what goes wrong and why (this talk)

Shape bias



Ritter, Samuel, et al. "Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study." ICML 2017.

Psychometrics

Scientific field focused on quantitative measurement practices

Focus is two-fold: (i) building instruments for measurement and (ii) development of theoretical approaches to measurement

Item response theory

Measure latent traits of test-takers and test questions ("items")

- GRE, GMAT
- Personality assessments
- EHR note comprehension

3 parameter logistic model (3PL)

$$p_{ij}(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

θ_j: latent ability of individual j

a_i: discriminatory ability of item i

b_i: difficulty of item i

c_i: guessing parameter for item i

3 parameter logistic model (3PL)

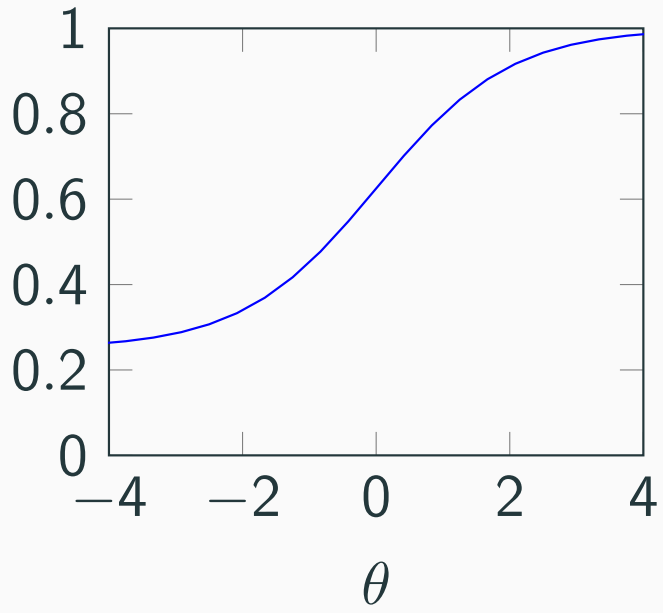
$$p_{ij}(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

θ_j: latent ability of individual j

a_i: discriminatory ability of item i

b_i: difficulty of item i

c_i: guessing parameter for item i



$$a_i = 1$$
$$b_i = 0$$
$$c_i = 0.25$$

3 parameter logistic model (3PL)

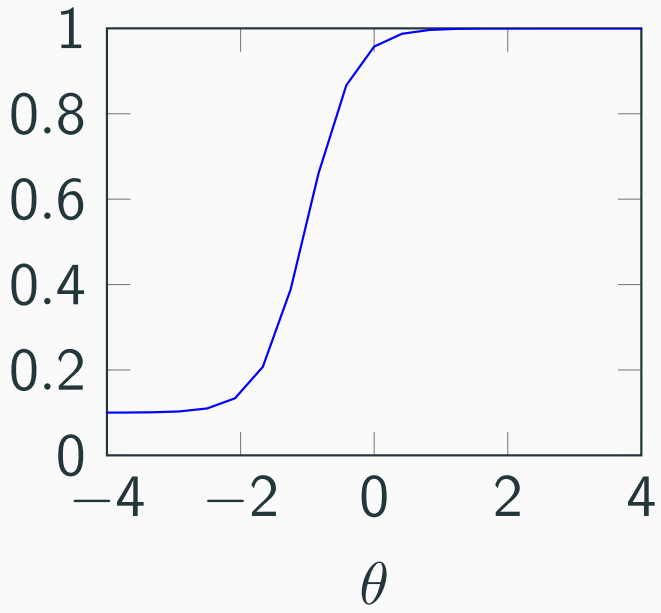
$$p_{ij}(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

θ_j: latent ability of individual j

a_i: discriminatory ability of item i

b_i: difficulty of item i

c_i: guessing parameter for item i



$$a_i = 3$$

$$b_i = -1$$

$$c_i = 0.1$$

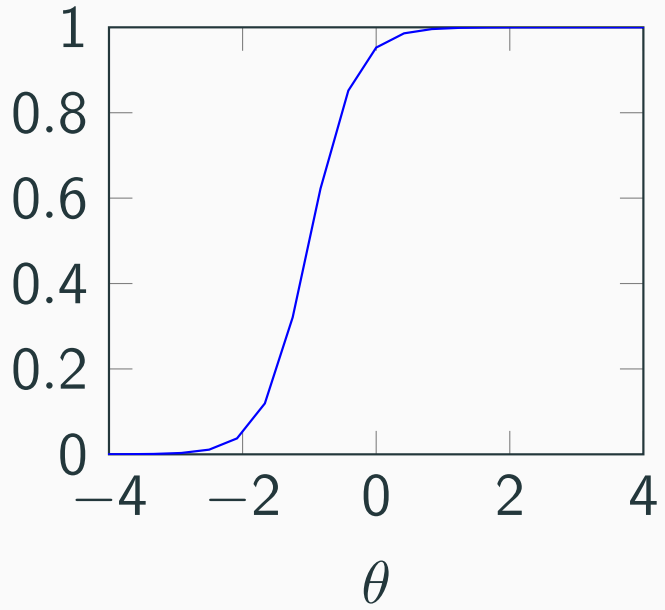
2 parameter logistic model (2PL)

$$p_{ij}(\theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}}$$

θ_j: latent ability of individual j

a_i: discriminatory ability of item i

b_i: difficulty of item i



$a_i = 3$

$b_i = -1$

Learning item parameters

$$p_{ij}(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

Chalmers, R. Philip. "mirt: A multidimensional item response theory package for the R environment." Journal of Statistical Software 48.6 (2012): 1-29.

Learning item parameters

$$p_{ij}(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

$$q_{ij}(\theta_j) = 1 - p_{ij}(\theta_j)$$

Chalmers, R. Philip. "mirt: A multidimensional item response theory package for the R environment." Journal of Statistical Software 48.6 (2012): 1-29.

Learning item parameters

$$p_{ij}(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

$$q_{ij}(\theta_j) = 1 - p_{ij}(\theta_j)$$

$$L = \prod_{j=1}^{J} \prod_{i=1}^{I} p_{ij}(\theta_j)^{y_{ij}} q_{ij}(\theta_j)^{(1-y_{ij})}$$

Chalmers, R. Philip. "mirt: A multidimensional item response theory package for the R environment." Journal of Statistical Software 48.6 (2012): 1-29.

Learning item parameters

$$p_{ij}(\theta_j) = c_i + \frac{1 - c_i}{1 + e^{-a_i(\theta_j - b_i)}}$$

$$q_{ij}(\theta_j) = 1 - p_{ij}(\theta_j)$$

$$L = \prod_{j=1}^{J} \prod_{i=1}^{I} p_{ij}(\theta_j)^{y_{ij}} q_{ij}(\theta_j)^{(1-y_{ij})}$$

Chalmers, R. Philip. "mirt: A multidimensional item response theory package for the R environment." Journal of Statistical Software 48.6 (2012): 1-29.

IRT for NLP

Gather human annotations for an existing dataset

IRT for NLP

Gather human annotations for an existing dataset

Fit an IRT model

IRT for NLP

Gather human annotations for an existing dataset

Fit an IRT model

Administer new “test” to DNN models

Original Dataset

Stanford Natural Language Inference (SNLI) corpus

550,000/10,000/10,000 train/dev/test sentence pairs

Recruit 1000 crowd workers to label sentence pairs

Premise	**Hypothesis**	**Label**
A little girl eating a sucker	A child eating candy	Entailment
People were watching the tournament in the stadium	The people are sitting outside on the grass	Contradiction
Two girls on a bridge dancing with the city skyline in the background	The girls are sisters.	Neutral
Nine men wearing tuxedos sing	Nine women wearing dresses sing	Contradiction

Raw score and IRT scores

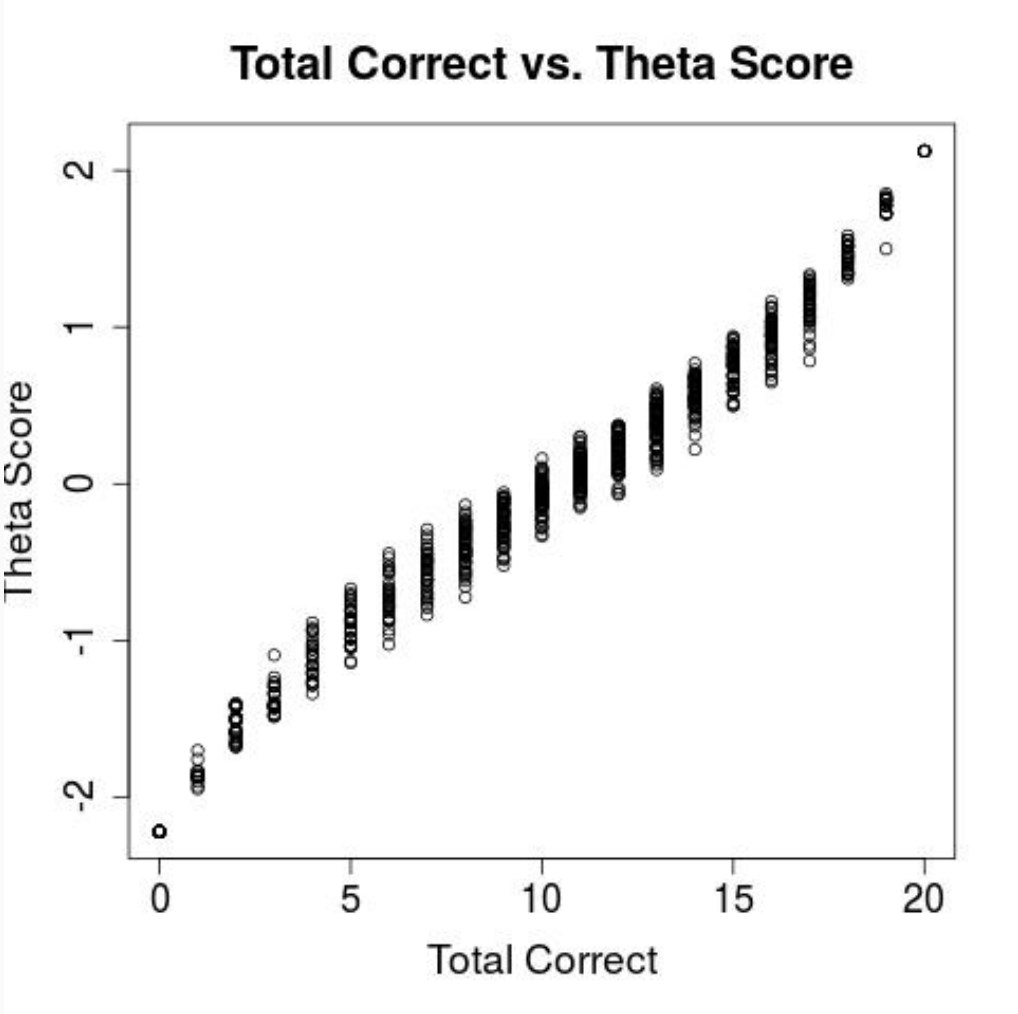
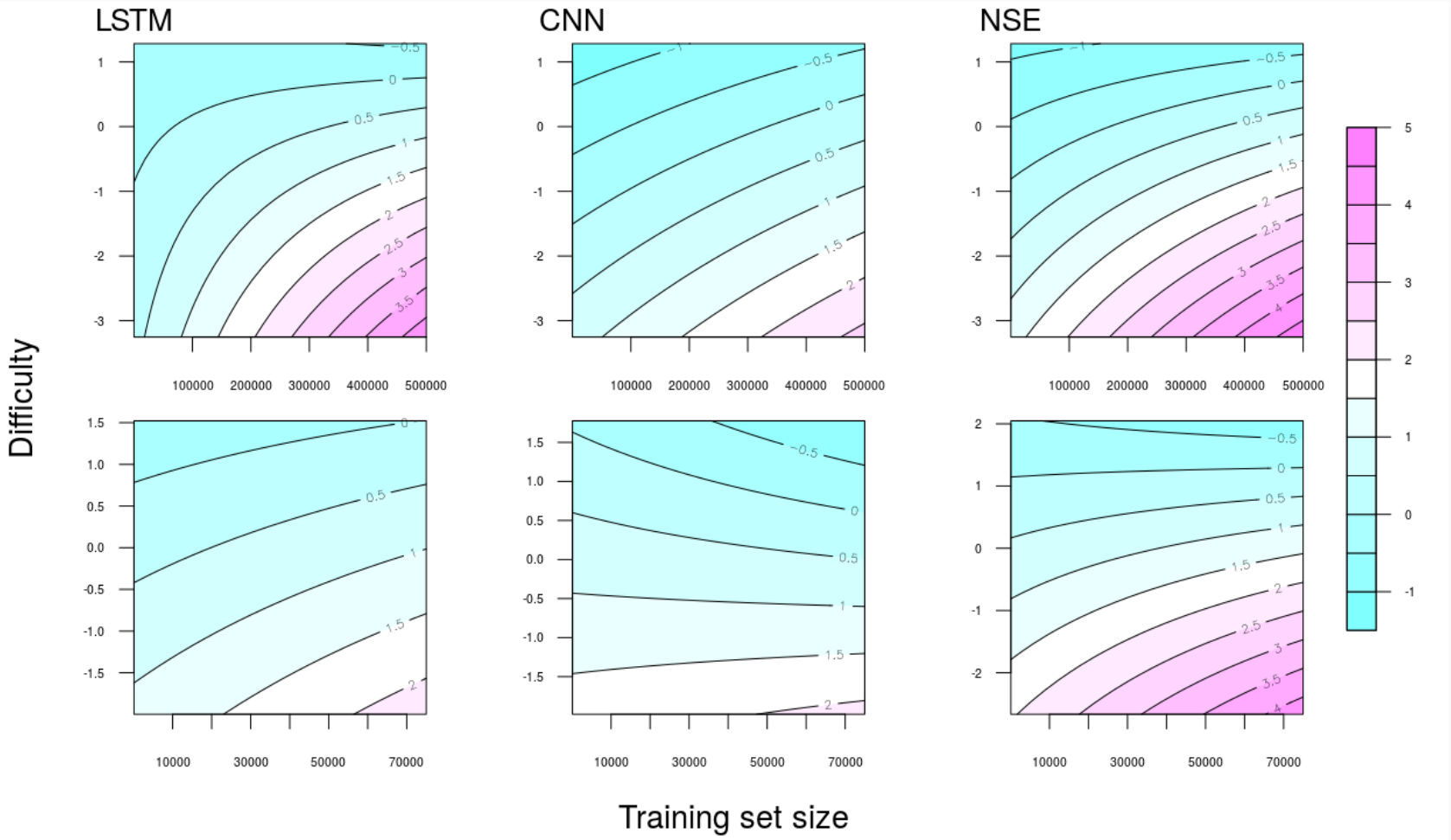


Figure 1: Plot of total correct answers vs. IRT scores.

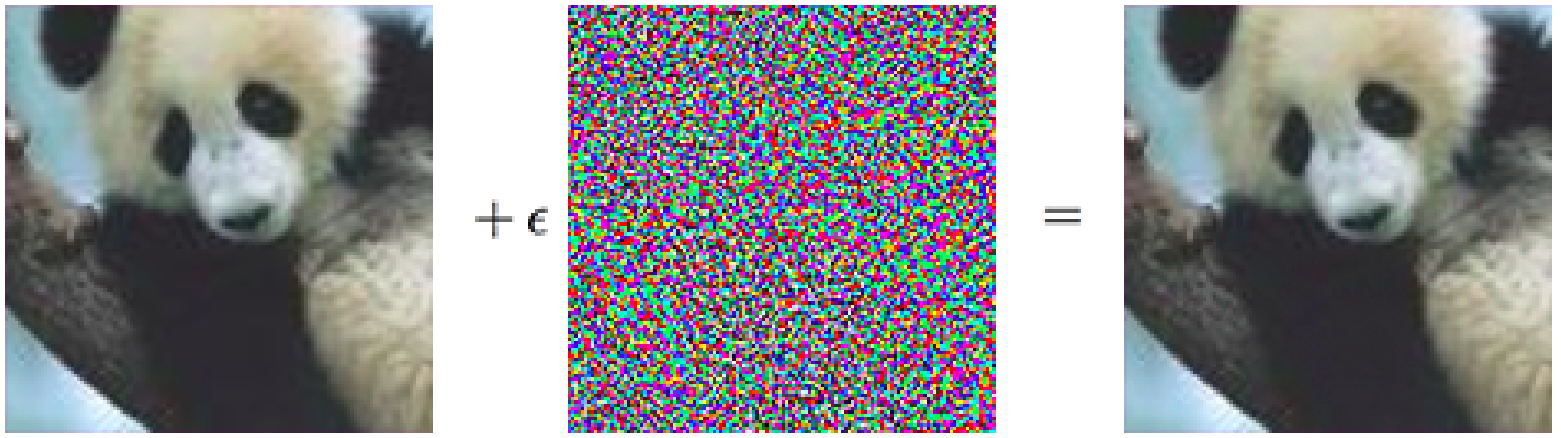
Evaluating DNN models

Item Set	Theta Score	Percentile	Test Acc.
5GS			
Entailment	-0.133	44.83%	96.5%
Contradiction	1.539	93.82%	87.9%
Neutral	0.423	66.28%	88%
4GS			
Contradiction	1.777	96.25%	78.9%
Neutral	0.441	67%	83%

Using learned difficulty

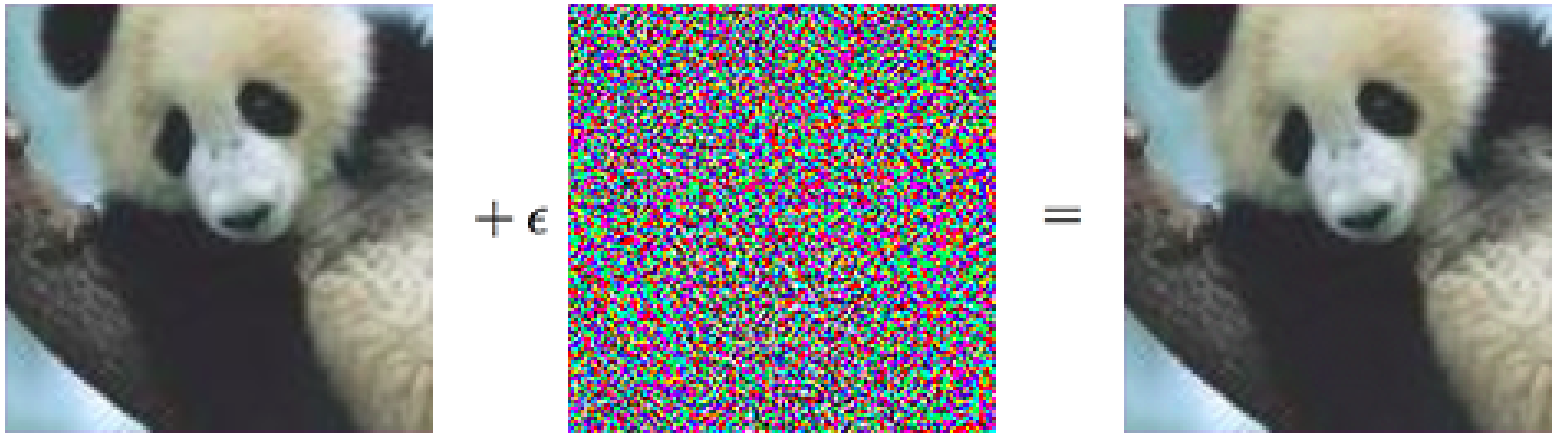


Motivation



Goodfellow, Ian J., et al. "Explaining and harnessing adversarial examples." ICLR 2015.

Motivation



Goodfellow, Ian J., et al. "Explaining and harnessing adversarial examples." ICLR 2015.

Fast Gradient Sign Method: Example

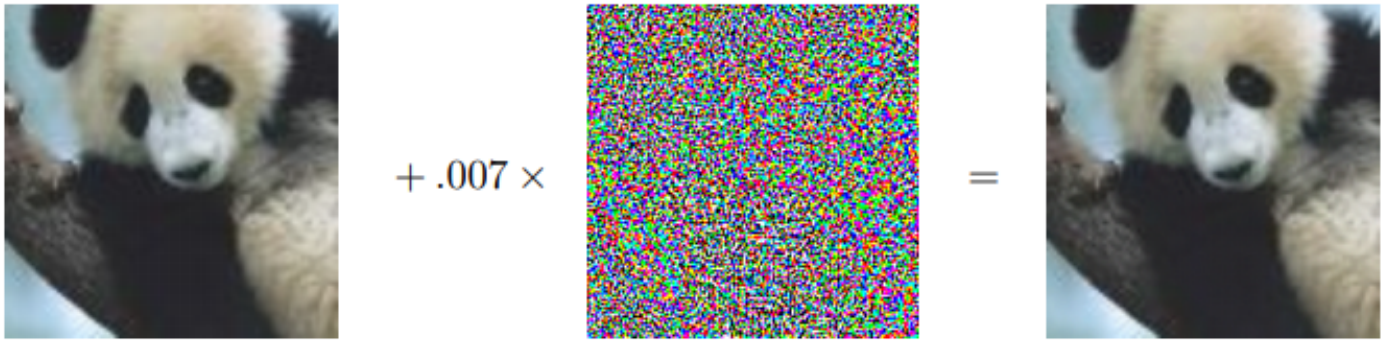


Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

Goodfellow, Ian J., et al. “Explaining and harnessing adversarial examples.” ICLR 2015.

Adversarial evaluation

How do you evaluate models susceptibility to adversarial attacks?

- Attack success rate
  - $f(x_{\text{adv}}) \neq y$
  - $f(x_{\text{adv}}) = T$ for target T
- Average distortion:

$$\Delta(X_{\text{adv}}, X) = \frac{1}{N} \sum_{i=1}^{N} ||(X_{\text{adv}})_i - X_i||_2$$

- Number of queries

Bhagoji, Arjun Nitin, et al. "Exploring the Space of Black-box Attacks on Deep Neural Networks." arXiv:1712.09491 (2017).

Table of contents

Traditional model evaluation

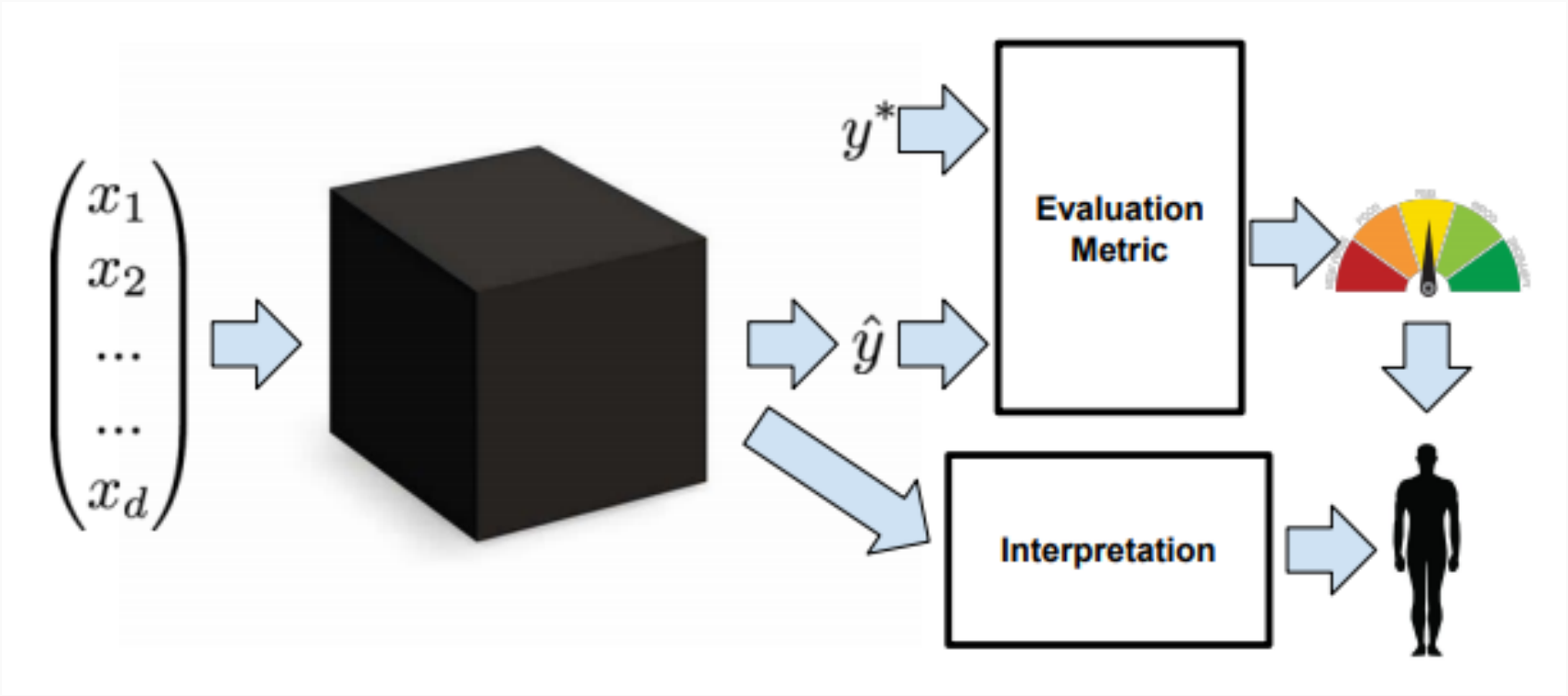
Hands-on: Evaluation

New evaluation methods

Interpretability

Hands-on: Interpretability

Evaluation and Interpretability



Lipton, Zachary C. "The Mythos of Model Interpretability." Queue 16.3 (2018): 30.

What do we mean by interpretability?

(1) Transparency and (2) Explainability ("post-hoc explanations[1]")

[1] Lipton, Zachary C. "The Mythos of Model Interpretability." Queue 16.3 (2018): 30.

What do we mean by interpretability?

(1) Transparency and (2) Explainability ("post-hoc explanations[1]")

Which do we want?

[1] Lipton, Zachary C. "The Mythos of Model Interpretability." Queue 16.3 (2018): 30.

What do we mean by interpretability?

(1) Transparency and (2) Explainability ("post-hoc explanations[1]")

Which do we want? When?

[1] Lipton, Zachary C. "The Mythos of Model Interpretability." Queue 16.3 (2018): 30.

What do we mean by interpretability?

(1) Transparency and (2) Explainability ("post-hoc explanations[1]")

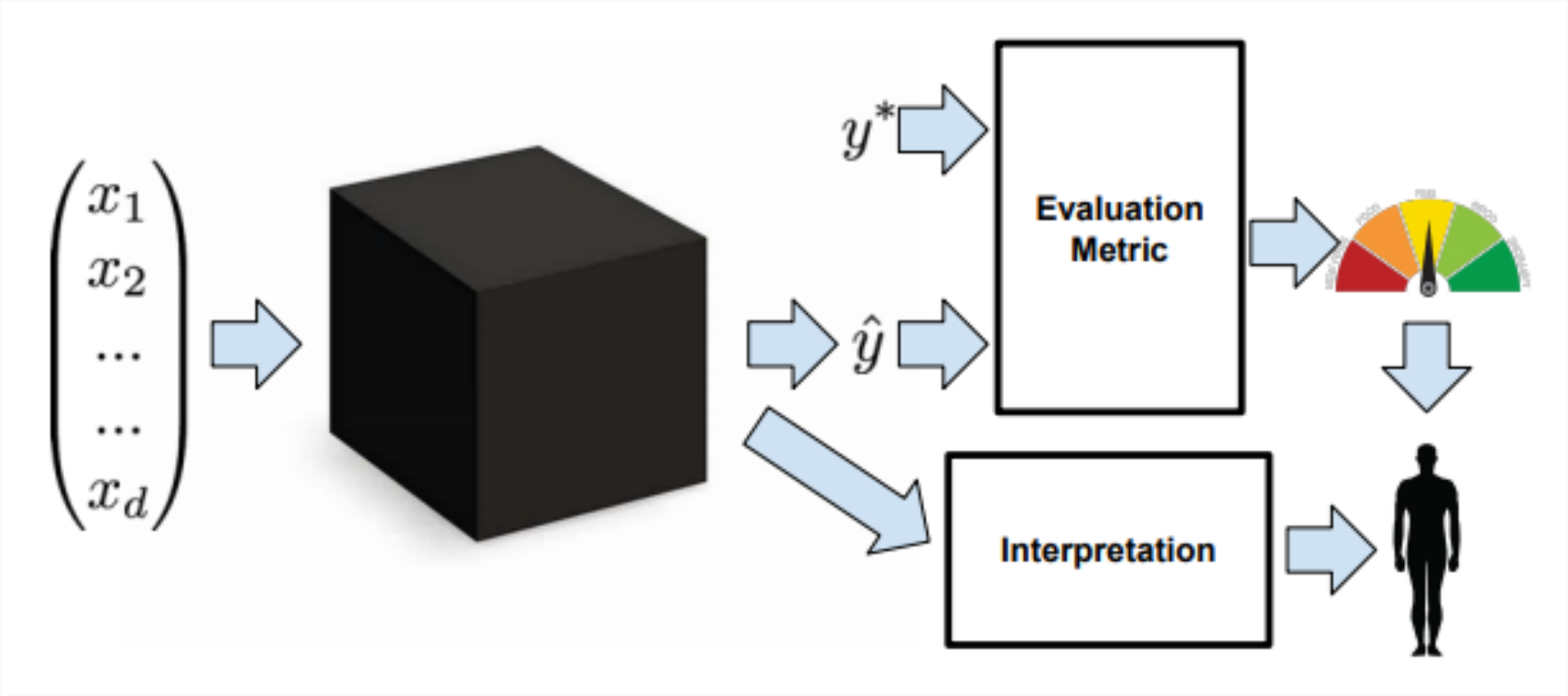
Which do we want? When? Why?

[1] Lipton, Zachary C. "The Mythos of Model Interpretability." Queue 16.3 (2018): 30.

Acknowledgment

Lipton, Zachary C. “The Mythos of Model Interpretability.” Queue 16.3 (2018): 30.

Evaluation-Interpretability Relationship



Lipton, Zachary C. "The Mythos of Model Interpretability." Queue 16.3 (2018): 30.

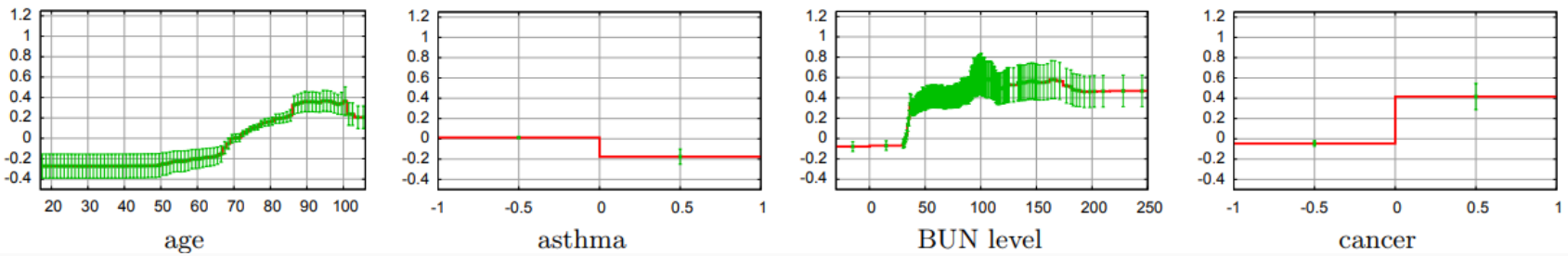
Why do we want interpretability?

Trust

https://waymo.com/ontheroad/

Why do we want interpretability?

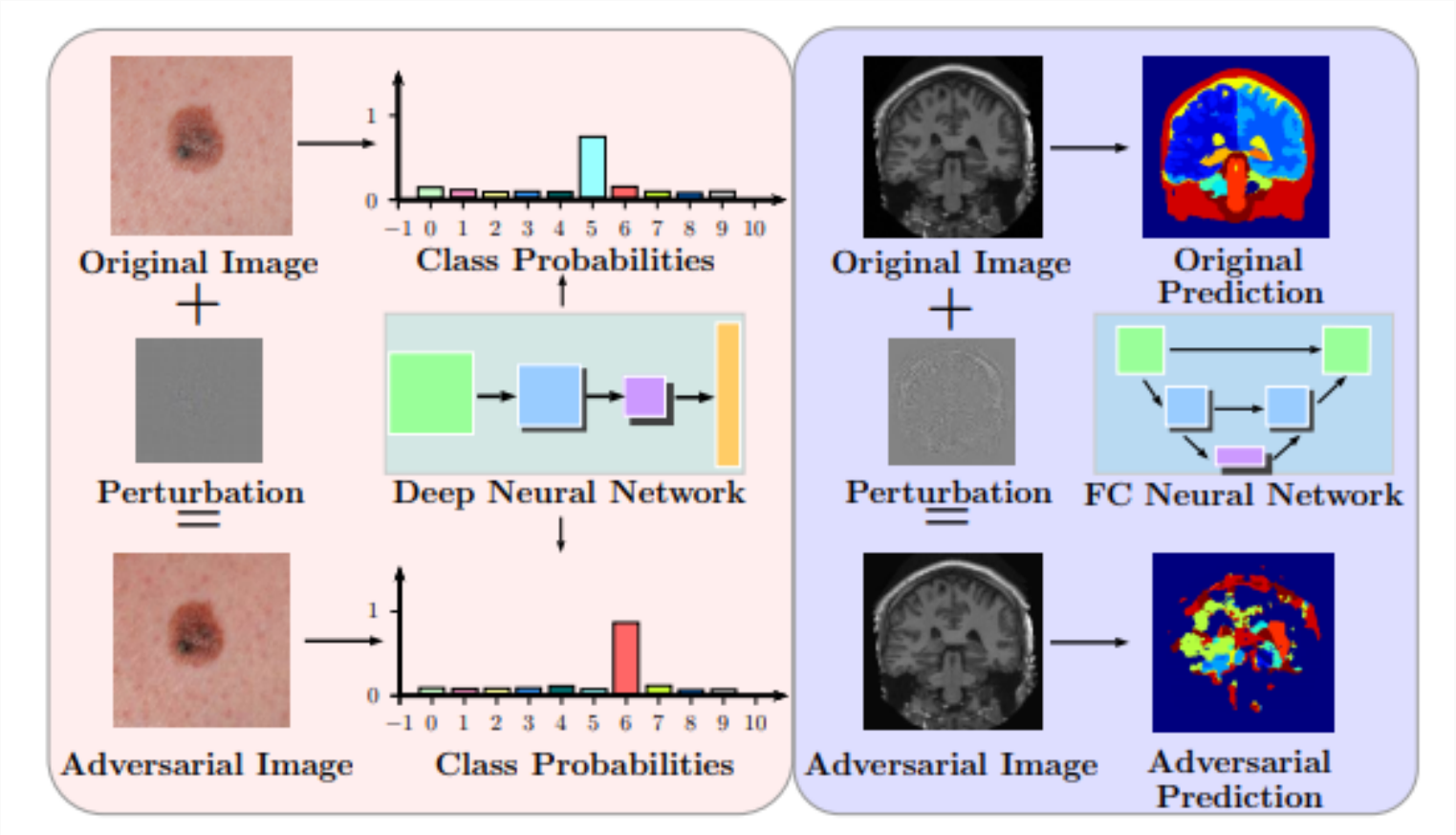
Causality



Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

Why do we want interpretability?

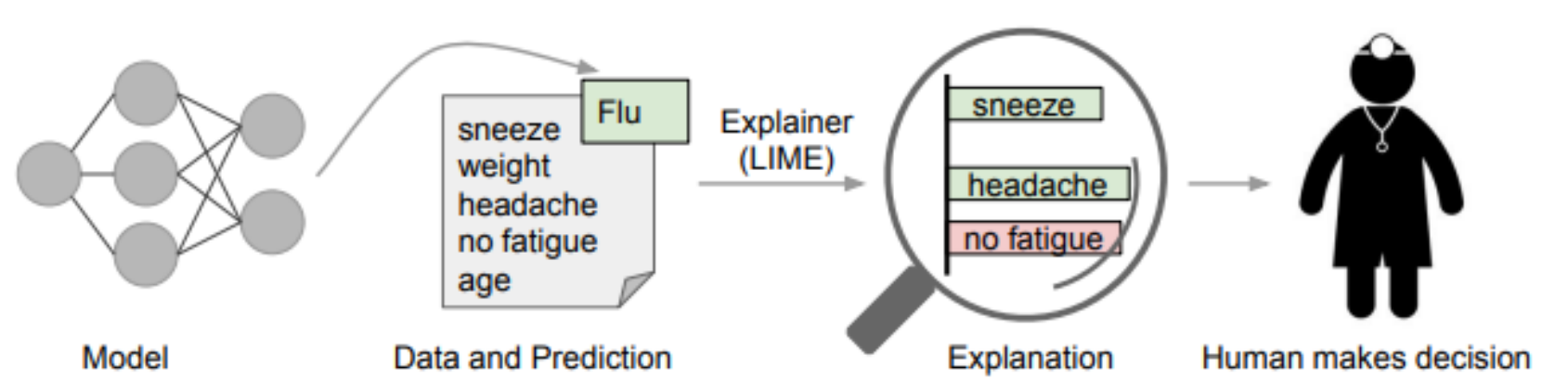
Transferability



Paschali, Magdalini, et al. "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging." arXiv:1804.00504 (2018).

Why do we want interpretability?

Informativeness

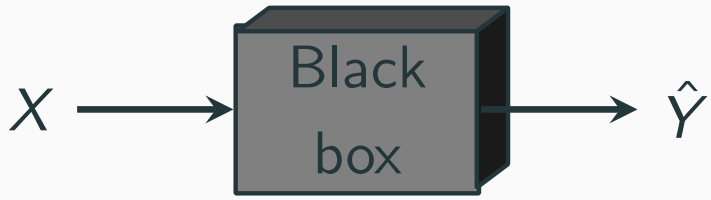


Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

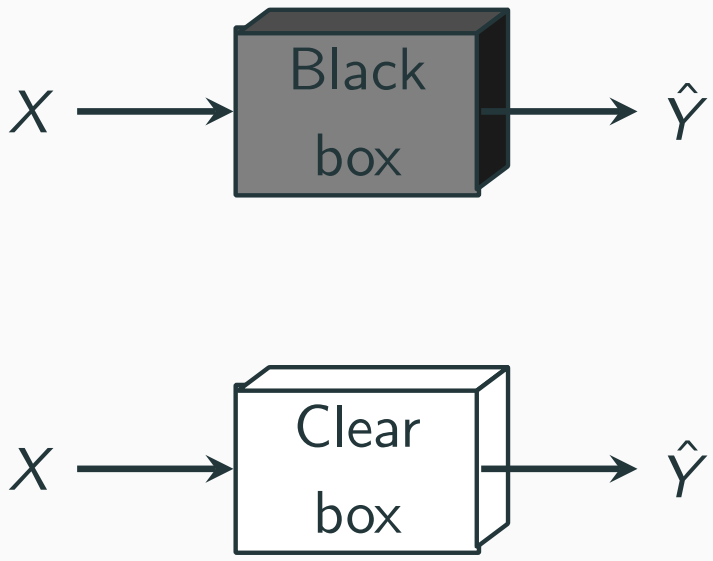
Why do we want interpretability?

Fairness (e.g. *the right to an explanation*)

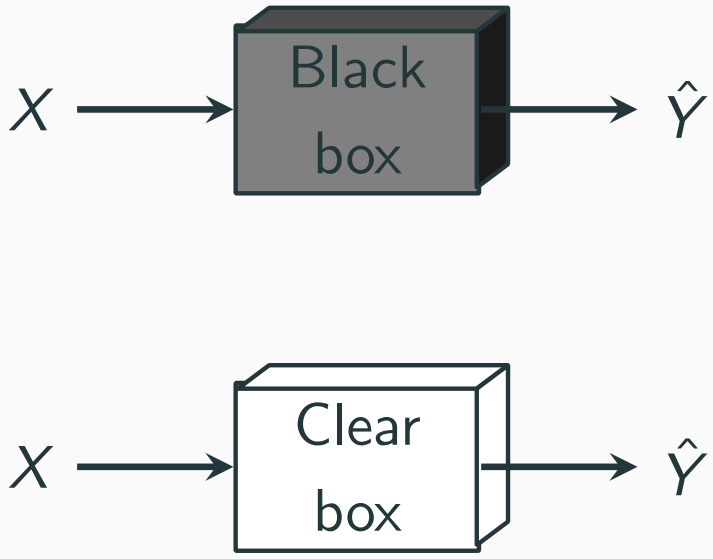
Transparency



Transparency



Transparency



Model transparency (simulatability)

Parameter transparency (decomposability)

Training transparency (algorithmic transparency)

Simulatability

Given the input data and model parameters:

Could a human produce an output?

Simulatability

Given the input data and model parameters:

- Could a human produce an output?
  - In some reasonable amount of time?

Simulatability

Given the input data and model parameters:

- Could a human produce an output?
  - In some reasonable amount of time?
- Example: BOW logistic regression

Simulatability

Given the input data and model parameters:

- Could a human produce an output?
  - In some reasonable amount of time?
- Example: BOW logistic regression
- Example: fully-connected DNN with hidden layer size 10

Decomposability

Each part of the model is intuitive

- Inputs
- Parameters
- Calculations

Decomposability

Each part of the model is intuitive

- Inputs
- Parameters
- Calculations

Ex.: Descriptive decision tree nodes

Decomposability

Each part of the model is intuitive

- Inputs
- Parameters
- Calculations

Ex.: Descriptive decision tree nodes

Ex.: Linear model parameters

Decomposability

Each part of the model is intuitive

- Inputs
- Parameters
- Calculations

Ex.: Descriptive decision tree nodes

Ex.: Linear model parameters

Caveat: Can be fragile depending on pre-processing

Algorithmic transparency

Insight into the decision-making process

Algorithmic transparency

Insight into the decision-making process

- Linear models?

Algorithmic transparency

Insight into the decision-making process

- Linear models?
- DNNs?

Algorithmic transparency

Insight into the decision-making process

- Linear models?
- DNNs?
- Humans?

Explanability

Human interpretability

After the fact interpretation

- Not part of model training

Text

a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a generous head that sustained life throughout . nothing out of the ordinary here , but a good brew still . body was kind of heavy , but not thick . the hop smell was excellent and enticing . very drinkable

---

very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . smells of coffee and roasted malt . has a major coffee-like taste with hints of chocolate . if you like black coffee , you will love this porter . creamy smooth mouthfeel and definitely gets smoother on the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

---

i really did not like this . it just seemed extremely watery . i dont ' think this had any carbonation whatsoever . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty nasty towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

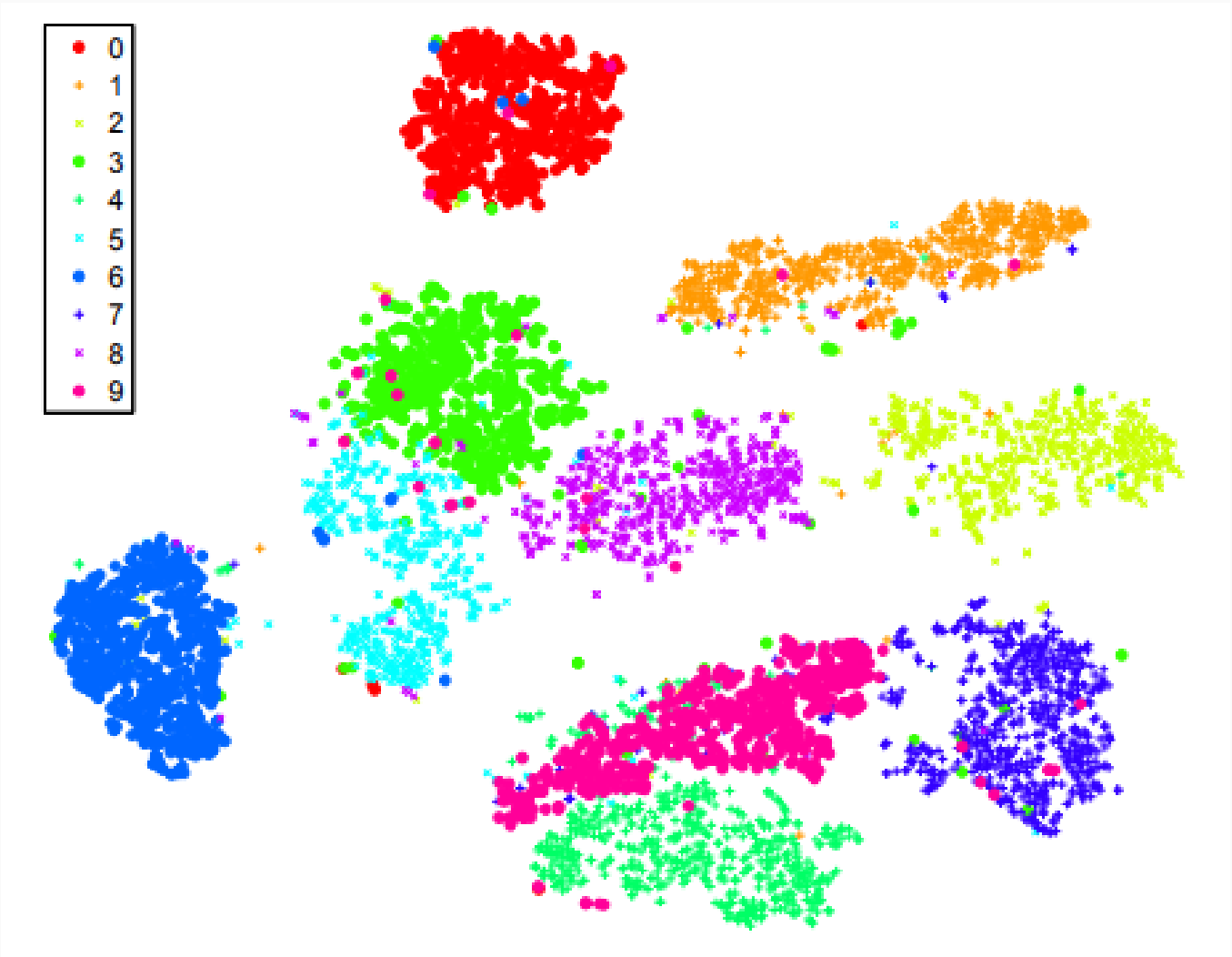
---

a : poured a nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around the glass , not too shabby . not terribly impressive though s : smells like a more guinness-y guinness really , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate ... ... m : relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation . not too shabby d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

Figure 3: Examples of extracted rationales indicating the sentiments of various aspects. The extracted texts for appearance, smell and palate are shown in red, blue and green color respectively. The last example is shortened for space.

Lei, Tao, Regina Barzilay, and Tommi Jaakkola. "Rationalizing Neural Predictions." EMNLP 2016.

Visualizations



Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.Nov (2008): 2579-2605.

Local explanations

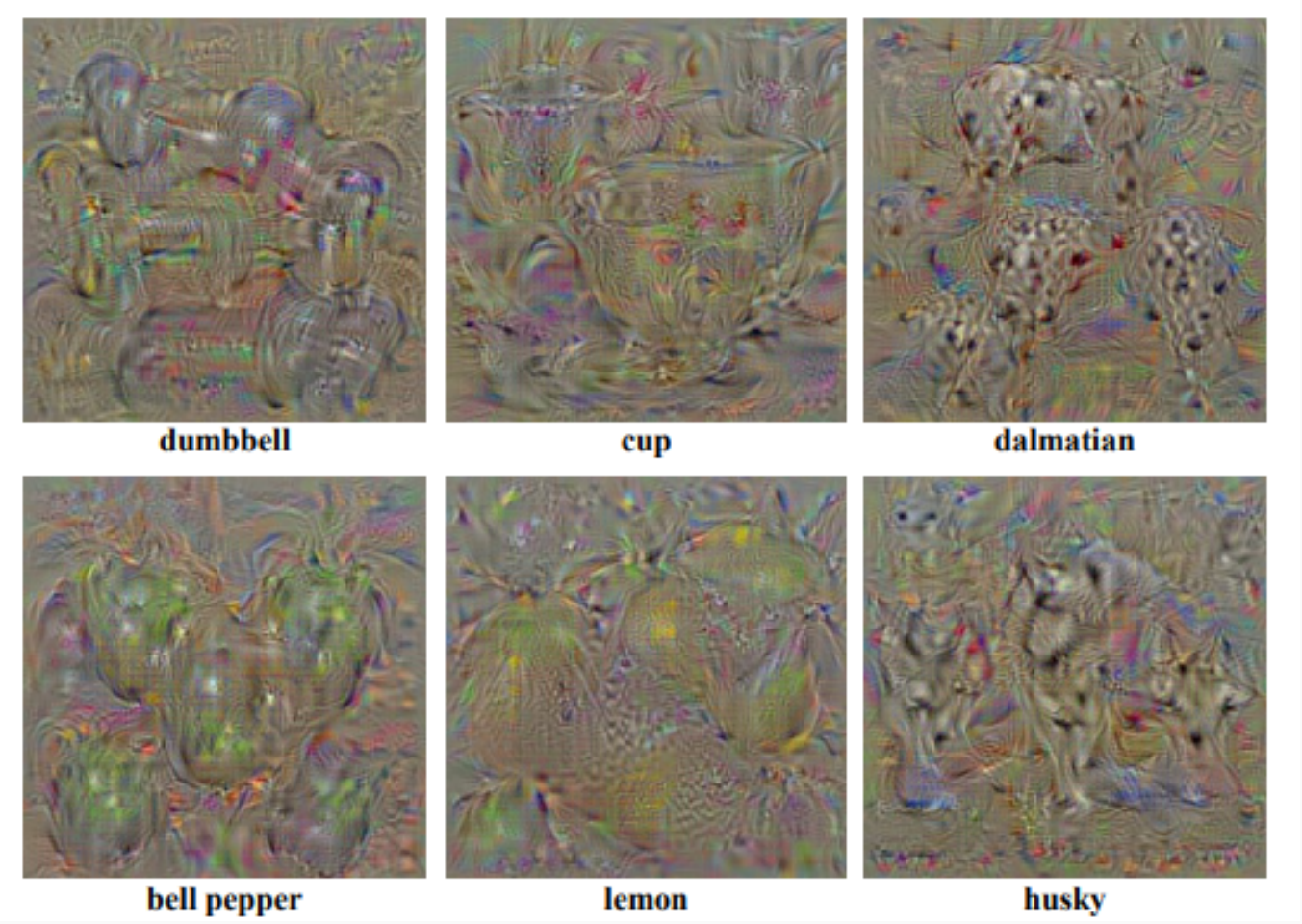
Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv:1312.6034 (2013).

Explanations by example

Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

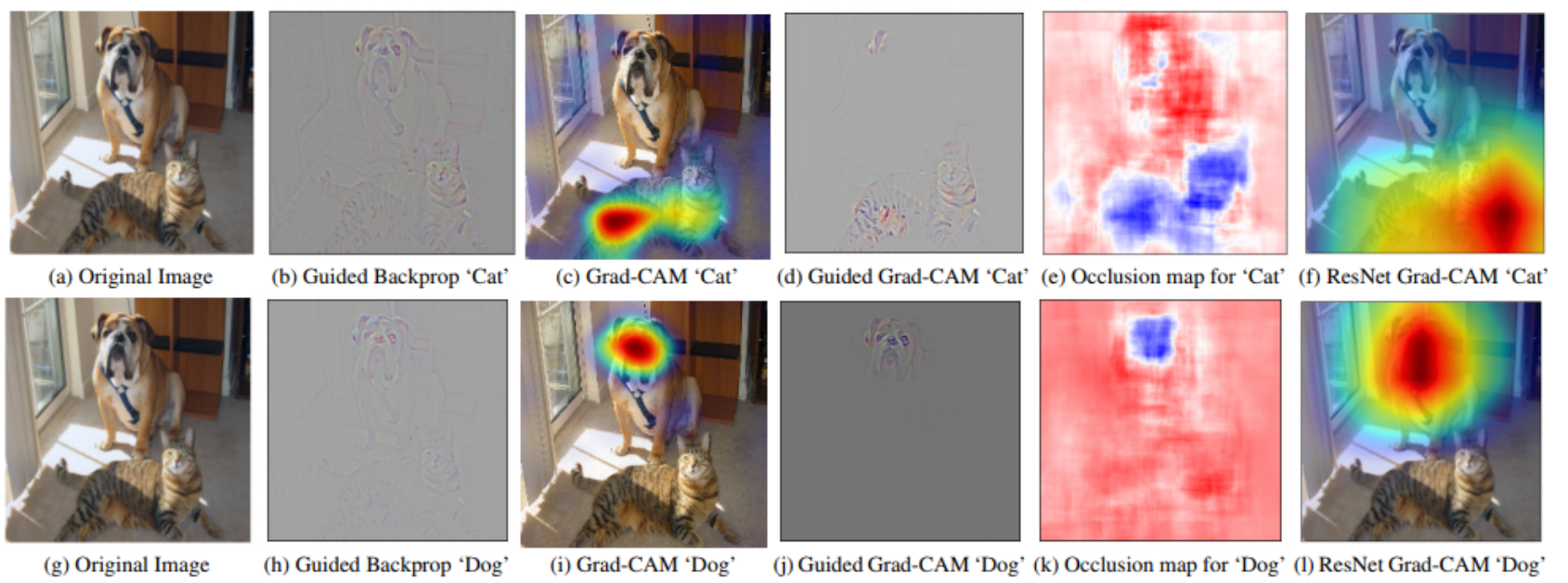
Saliency maps



$$\arg\max_I S_c(I) - \lambda||I||_2^2$$

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv:1312.6034 (2013).

Gradient based localization: Grad-CAM

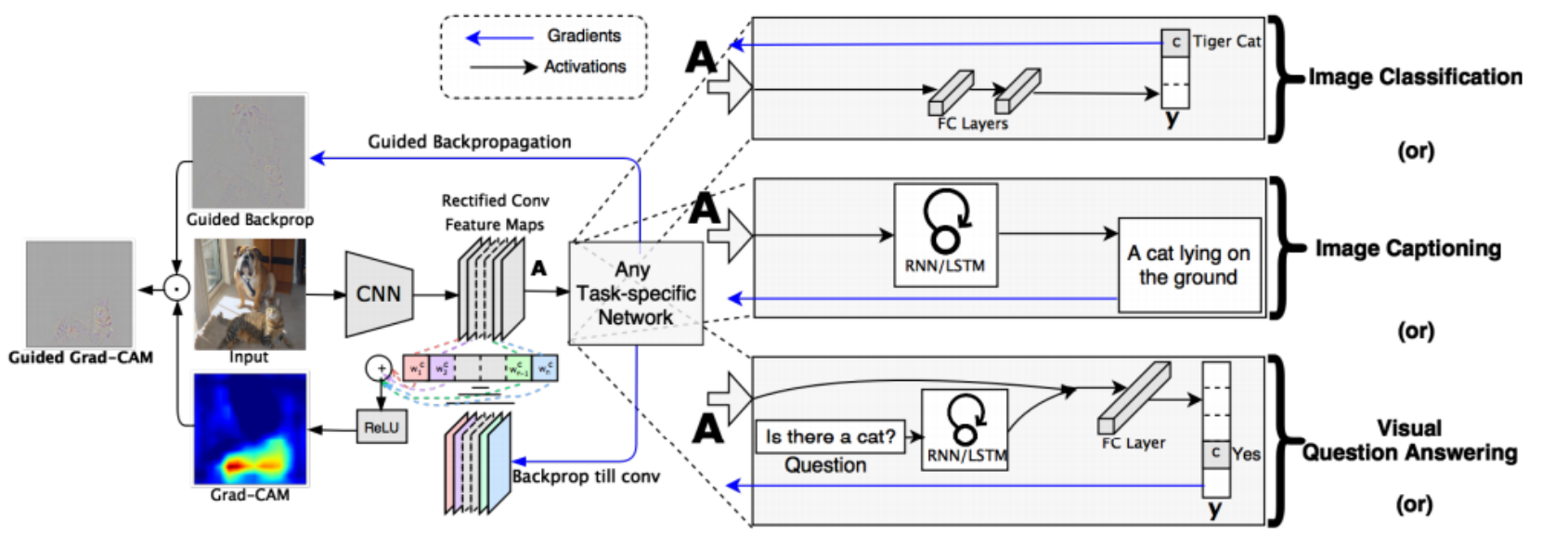


(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat' (d) Guided Grad-CAM 'Cat' (e) Occlusion map for 'Cat' (f) ResNet Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog' (j) Guided Grad-CAM 'Dog' (k) Occlusion map for 'Dog' (l) ResNet Grad-CAM 'Dog'

Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." ICCV. 2017.

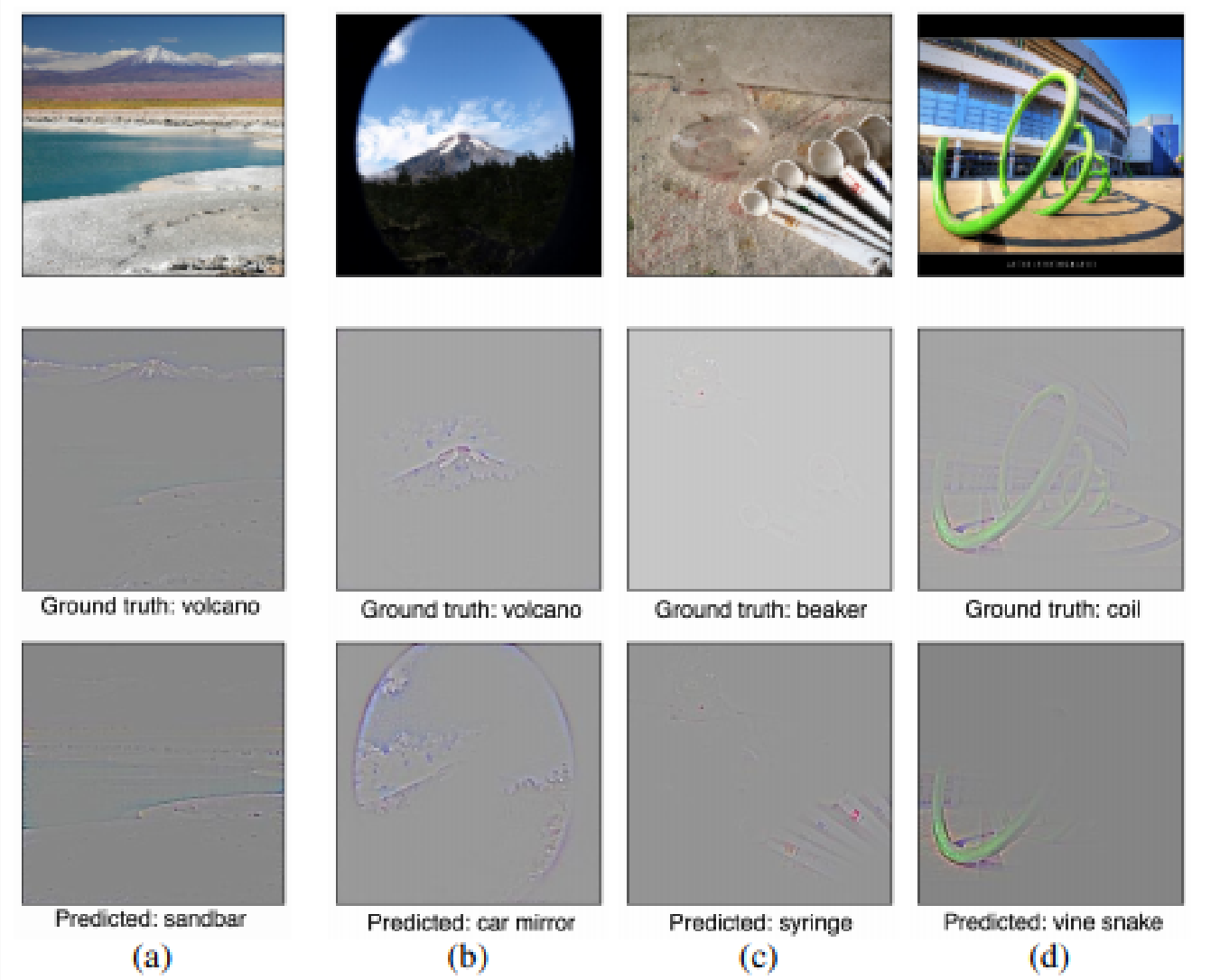
Gradient based localization: Grad-CAM



Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." ICCV. 2017.

Gradient based localization: Grad-CAM

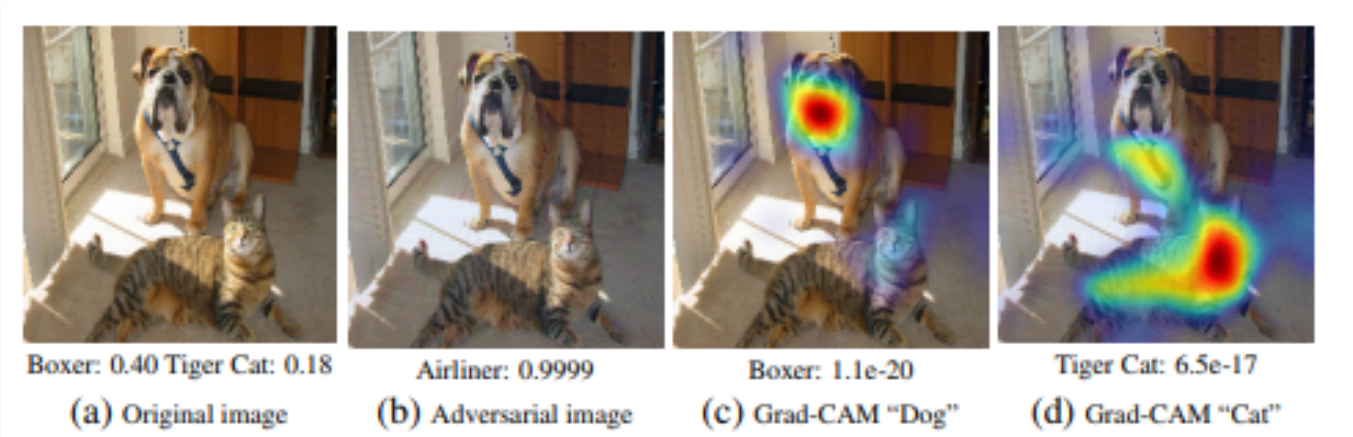
Analyzing failures



Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based

Gradient based localization: Grad-CAM

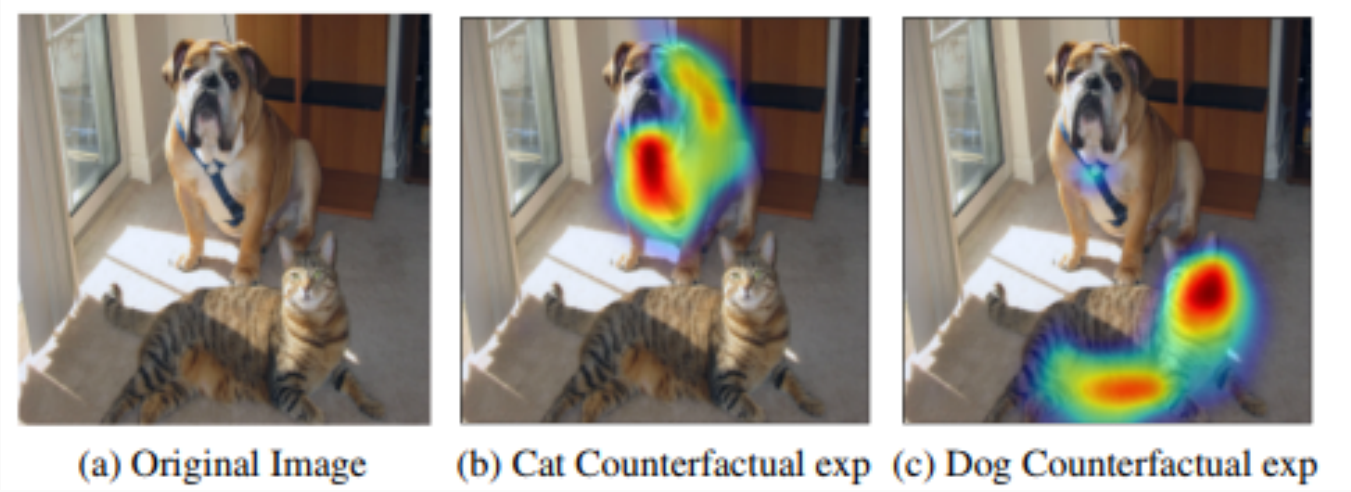
Handling adversarial noise



Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." ICCV. 2017.

Gradient based localization: Grad-CAM

Counterfactuals



(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

---

Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." ICCV. 2017.

Interpretability takeaways

Linear models are not strictly more interpretable than DNNs

Interpretability takeaways

Linear models are not strictly more interpretable than DNNs

Claims about interpretability must be qualified

Interpretability takeaways

Linear models are not strictly more interpretable than DNNs

Claims about interpretability must be qualified

Transparency *may* be at odds with end goal

Interpretability takeaways

Linear models are not strictly more interpretable than DNNs

Claims about interpretability must be qualified

Transparency *may* be at odds with end goal

Misleading post-hoc interpretations

Conclusion

Evaluation and interpretability are linked

Conclusion

Evaluation and interpretability are linked

Always know why you want a certain evaluation/visualization/etc.

Conclusion

Evaluation and interpretability are linked

Always know why you want a certain evaluation/visualization/etc.

Consider stakeholders, expectations, and ultimate goals

Thank you!

email: lalor@cs.umass.edu

web: http://jplalor.github.io

Table of contents

Traditional model evaluation

Hands-on: Evaluation

New evaluation methods

Interpretability

Hands-on: Interpretability

Building blocks of interpretability

https://distill.pub/2018/building-blocks/

AMIA is the professional home for more than 5,400 informatics professionals, representing frontline clinicians, researchers, public health experts and educators who bring meaning to data, manage information and generate new knowledge across the research and healthcare enterprise.

INFORMATICS PROFESSIONALS. LEADING THE WAY.

@AMIAInformatics

@AMIAinformatics

Official Group of AMIA

@AMIAInformatics

#WhyInformatics