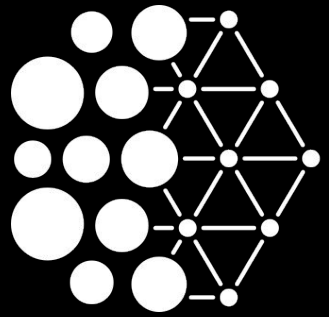
Quebec
Artificial
Intelligence
Institute



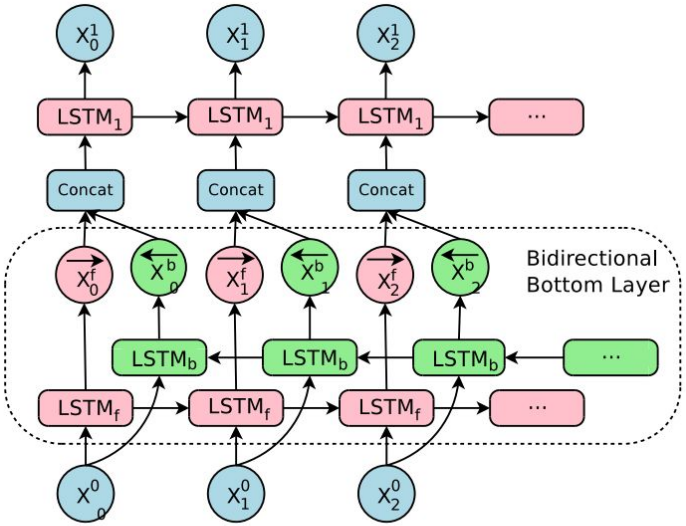
Mila

Robustness in machine learning
Week 1 - Part 1

Gaétan Marceau Caron
gaetan.marceau.caron@mila.quebec

Machine learning generates softwares!

Google replaced **500K lines** of code for automatic translation by a **500 lines of code** with **much better performances**.



[1] Systems and Software for Machine Learning at Scale with Jeff Dean at 27:00, twiML.
[2] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv:1609.08144. 2016.

Where are the requirements?

The requirements are encoded in:

1. the training dataset,
2. the training loss,
3. the architecture.

} feedback for training the model

Where are the requirements?

The requirements are encoded in:

1. the training dataset,
2. the training loss,
3. the architecture.

} feedback for training the model

What are the guarantees that the learned program respects the requirements?

Where are the requirements?

The requirements are encoded in:

1. the training dataset,
2. the training loss,
3. the architecture.

} feedback for training the model

What are the guarantees that the learned program respects the requirements?

We only have statistical guarantees based on the *identically and independently distributed assumption (iid)*.

Where are the requirements?

The requirements are encoded in:

1. the training dataset,
2. the training loss,
3. the architecture.

} feedback for training the model

What are the guarantees that the learned program respects the requirements?

This is not sufficient for real-world applications!

We only ntees based on the *identically and indep assumption (iid)*.

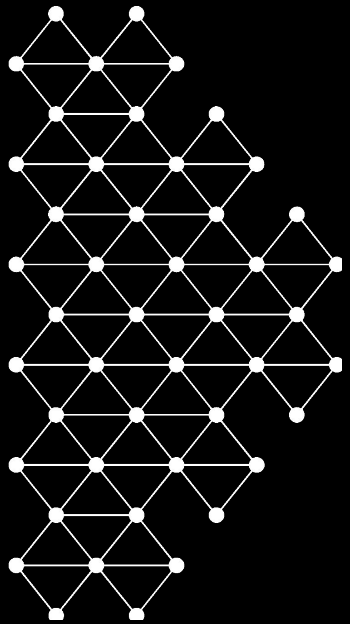
Topics that we will cover today

1. Limitations of *our* learning framework

2. Improving the evaluation protocol

Quebec
Artificial
Intelligence
Institute

Mila



Limitations of our learning framework

Definition

robustness. The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions. *See also:* **error tolerance; fault tolerance.**

IEEE Standards Coordinating Committee. "IEEE Standard Glossary of Software Engineering Terminology (IEEE Std 610.12-1990). Los Alamitos." CA: IEEE Computer Society 169. 1990.

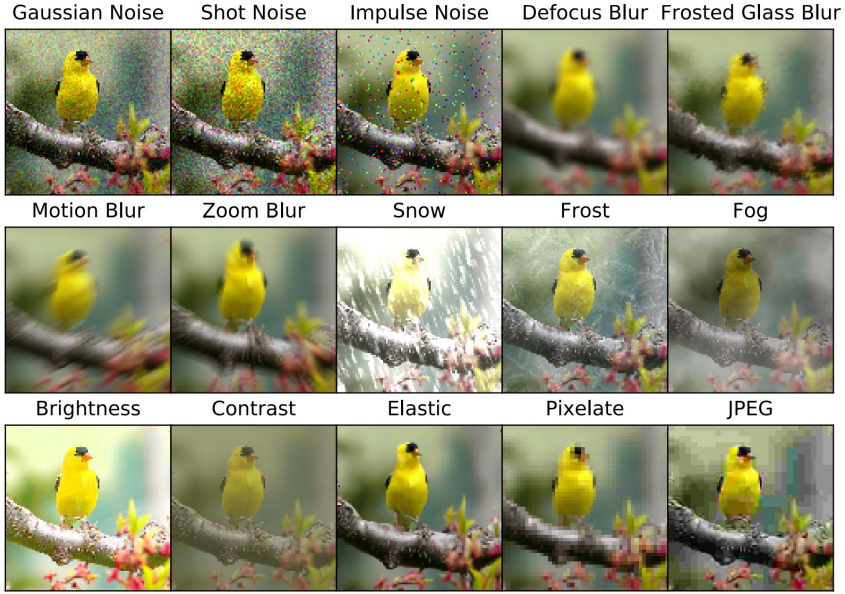
Incomplete definition

robustness. The degree to which a system or component can function correctly in the presence of invalid inputs or stressful environmental conditions. *See also:* **error tolerance; fault tolerance.**

IEEE Standards Coordinating Committee. "IEEE Standard Glossary of Software Engineering Terminology (IEEE Std 610.12-1990). Los Alamitos." CA: IEEE Computer Society 169. 1990.

Limitations of the IID hypothesis

Stating that examples are *independent and identically distributed* is a ***judgment call***.



Dan Hendrycks, Thomas G. Dietterich: Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. ICLR (Poster) 2019

Limitations of the IID hypothesis

Stating that examples are *independent and identically distributed* is a ***judgment call***.

$\boldsymbol{x}$

"panda"
57.7% confidence

$\boldsymbol{x} +$
$\epsilon \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy: Explaining and Harnessing Adversarial Examples. ICLR (Poster) 2015

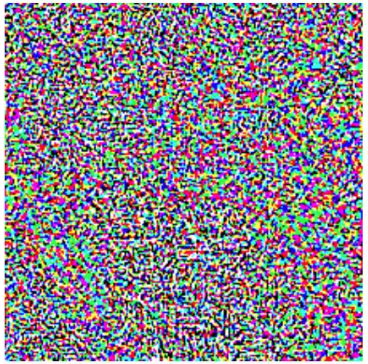
Limitations of the IID hypothesis

Stating that examples are *independent and identically distributed* is a ***judgment call***.

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

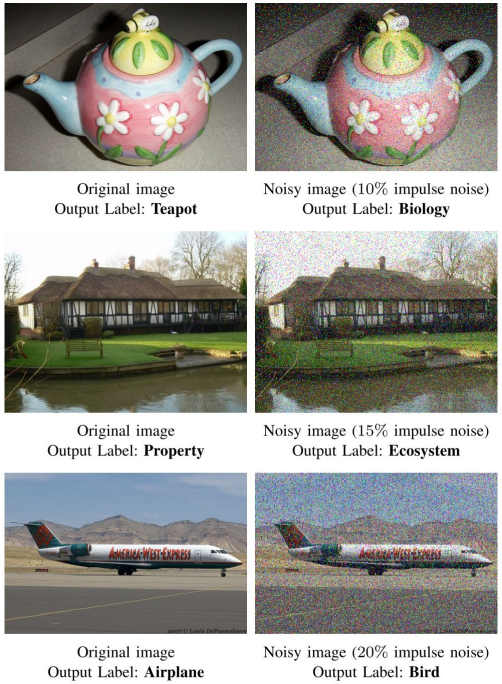
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy: Explaining and Harnessing Adversarial Examples. ICLR (Poster) 2015

Google’s Cloud Vision Is Not Robust To Noise

“Adding [...] noise is enough to deceive the API.” (July, 2017)



Hosseini, Hossein, Baicen Xiao, and Radha Poovendran. "Google's cloud vision api is not robust to noise." 16th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2017.

Out-of-distribution

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

CNN are biased towards texture

(a) Texture image

81.4%	**Indian elephant**
10.3%	indri
8.2%	black swan

(b) Content image

71.1%	**tabby cat**
17.3%	grey fox
3.3%	Siamese cat

(c) Texture-shape cue conflict

63.9%	**Indian elephant**
26.4%	indri
9.6%	black swan

Figure 1: Classification of a standard ResNet-50 of **(a)** a texture image (elephant skin: only texture cues); **(b)** a normal image of a cat (with both shape and texture cues), and **(c)** an image with a texture-shape cue conflict, generated by style transfer between the first two images.

Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." arXiv:1811.12231. 2018.

NLP models are brittle and spurious [1]

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

"In this adversarial setting, the accuracy of **sixteen published models** drops from an average of **75% F1 score to 36%**; when the adversary is allowed to add **ungrammatical sequences of words**, average accuracy on four models **decreases further to 7%**." [2]

[1] Ana Marasović, "NLP's generalization problem, and how researchers are tackling it", The Gradient, 2018.
[2] Robin Jia, Percy Liang: Adversarial Examples for Evaluating Reading Comprehension Systems. EMNLP 2017: 2021-2031

NLP models are brittle and spurious

- The metrics are flawed (BLEU, ROUGE, ...)

[1] Tom McCoy, Ellie Pavlick, Tal Linzen: Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL 2019
[2] de Vries, Harm, Dzmitry Bahdanau, and Christopher Manning. "Towards ecologically valid research on language user interfaces." arXiv:2007.14435 2020

NLP models are brittle and spurious

- The metrics are flawed (BLEU, ROUGE, ...)
- The datasets are too limited
  - Heuristics can work pretty well [1]

[1] Tom McCoy, Ellie Pavlick, Tal Linzen: Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL 2019
[2] de Vries, Harm, Dzmitry Bahdanau, and Christopher Manning. "Towards ecologically valid research on language user interfaces." arXiv:2007.14435 2020

NLP models are brittle and spurious

- The metrics are flawed (BLEU, ROUGE, ...)
- The datasets are too limited
  - Heuristics can work pretty well [1]
  - Not ecologically valid research [2]

Deviation	Project
Synthetic language	BabyAI (Chevalier-Boisvert et al., 2019) CLEVR (Johnson et al., 2017) CFQ (Keysers et al., 2019) GQA (Hudson and Manning, 2019)
Artificial task	GuessWhat (De Vries et al., 2017) CerealBar (Suhr et al., 2019) CoDraw (Kim et al., 2019) VisionAndLanguage (Anderson et al., 2018)
Not working with prospective users	Visual Question Answering (Antol et al., 2015) Visual Dialog (Das et al., 2017) Spider (Yu et al., 2018) SQuAD (Rajpurkar et al., 2016)
Scripts and priming	MultiWOZ (Budzianowski et al., 2018) ALFRED (Shridhar et al., 2020) CoSQL (Yu et al., 2019a) Sparc (Yu et al., 2019b) AirDialogue (Wei et al., 2018) Overnight (Wang et al., 2015)
Single-turn interfaces	Advising (Finegan-Dollak et al., 2018) MS Marco (Bajaj et al., 2016) Natural Questions (Kwiatkowski et al., 2019) DuReader (He et al., 2018)

Table 7: Five common deviations from the proposed ecologically valid research procedure. For each deviation we list a number of recent LUI benchmarks that suffer from it.

[1] Tom McCoy, Ellie Pavlick, Tal Linzen: Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL. 2019
[2] de Vries, Harm, Dzmitry Bahdanau, and Christopher Manning. "Towards ecologically valid research on language user interfaces." arXiv:2007.14435. 2020

NLP models are brittle and spurious

- The metrics are flawed (BLEU, ROUGE, ...)
- The datasets are too limited
  - Heuristics can work pretty well [1]
  - Not ecologically valid research [2]
- Evaluation protocol does not reflect deployment environment [2]

[1] Tom McCoy, Ellie Pavlick, Tal Linzen: Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL. 2019
[2] de Vries, Harm, Dzmitry Bahdanau, and Christopher Manning. "Towards ecologically valid research on language user interfaces." arXiv:2007.14435. 2020

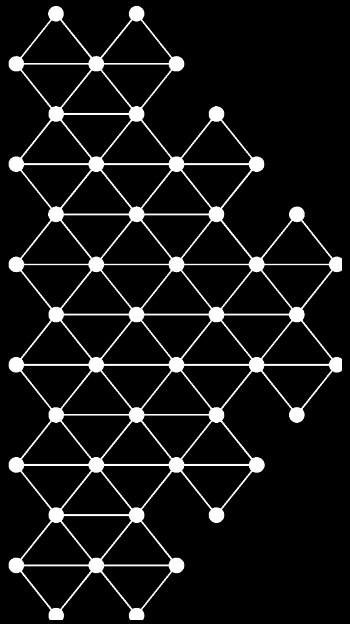
Lazy programmer

"I choose a lazy person to do a hard job.
Because a lazy person will find
an easy way to do it."

Bill Gates

Quebec
Artificial
Intelligence
Institute

Mila



Building trustworthy AI

Improving the evaluation protocol

1. Implement the evaluation protocol described in the MOOC.

2. Create groups of data to verify robustness to new factors like different:
   - measuring circumstances (camera types),
   - locations (Montreal vs Paris),
   - periods of time (day vs night),
   - demographic or phenotypic attributes (see module on Fairness), ...

3. Test your model with *different metrics to identify its limitations.*

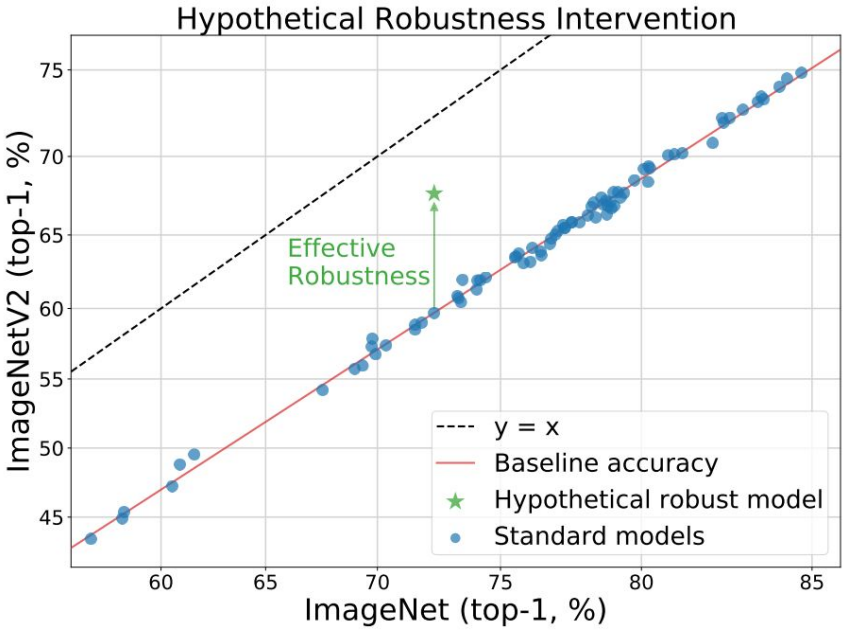
Improving the performance reporting

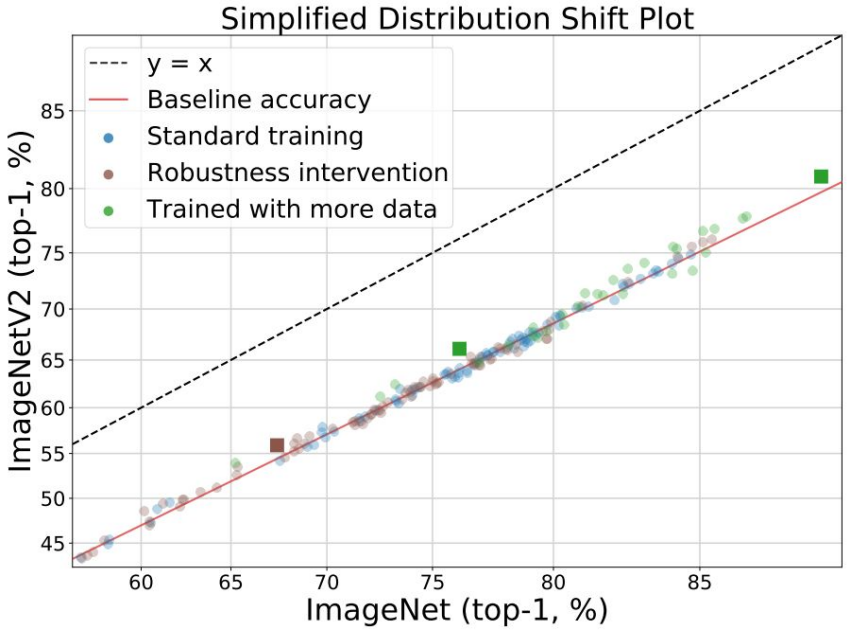
Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru: Model Cards for Model Reporting. FAT 2019: 220-229

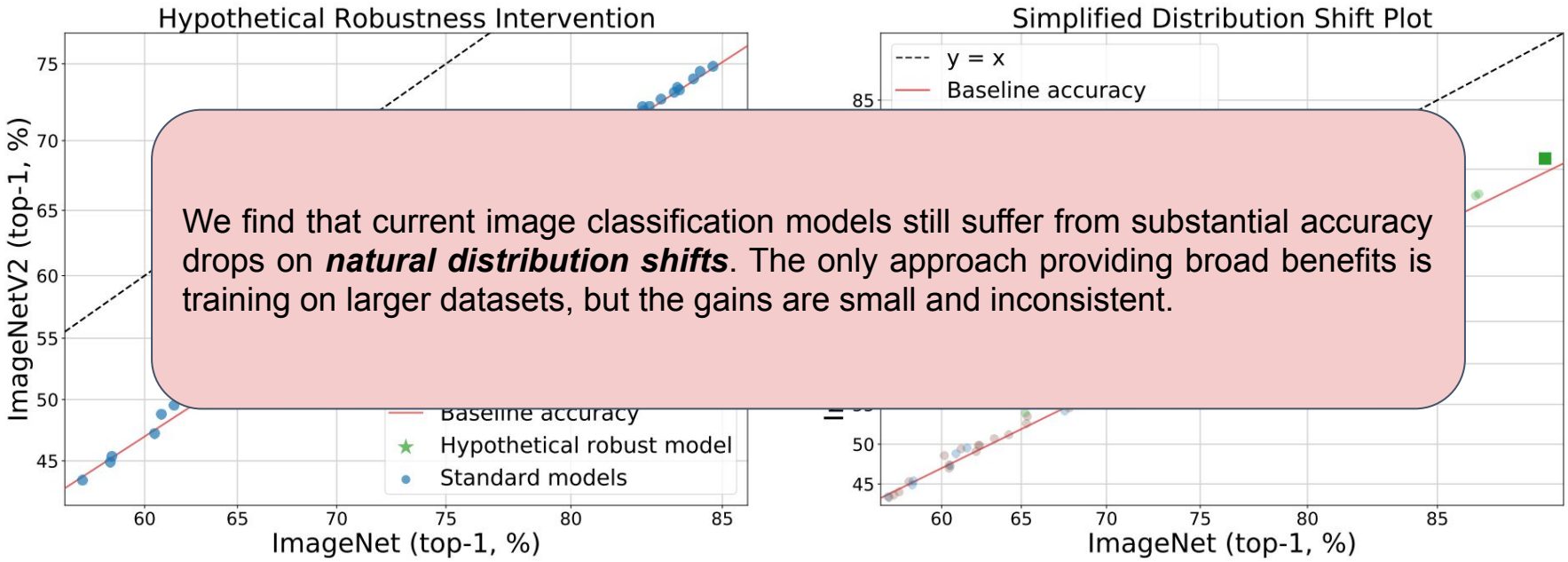
Robustification techniques





Taori, Rohan, et al. "Measuring robustness to natural distribution shifts in image classification." Advances in Neural Information Processing Systems 33 (2020).

Robustification techniques



We find that current image classification models still suffer from substantial accuracy drops on ***natural distribution shifts***. The only approach providing broad benefits is training on larger datasets, but the gains are small and inconsistent.

Taori, Rohan, et al. "Measuring robustness to natural distribution shifts in image classification." Advances in Neural Information Processing Systems 33 (2020).

Discussion

- While models are deployed in real-world applications, we discover their limitations.

- Many models are not robust: “they are solving datasets, not tasks”.

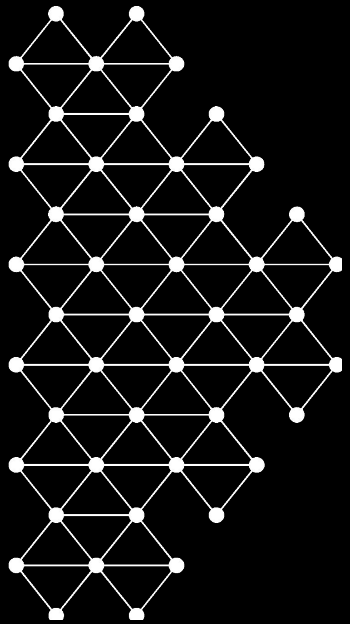
- Yet, better to have “lazy programmers” to solve your task than no programmers at all.

- Robustness is an active research topic, which shows that DL is maturing.

- Testing the limits of models is part of the job of AI developers.

Quebec
Artificial
Intelligence
Institute

Mila



Questions?

Robustification techniques

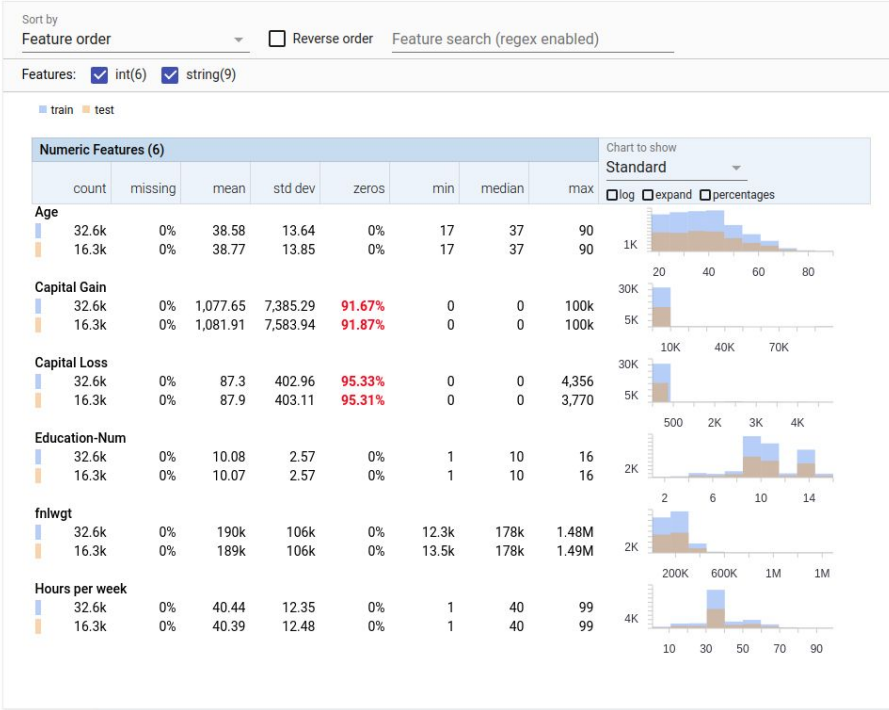
- Data pre-processing and data augmentation
- Inductive bias
- Transfer learning
  - domain adaptation
  - supervised pre-training
  - self-supervised pre-training (next week!)
- Model calibration
- and more...

Source: Katie Drazdauskaite - Unsplash

Djolonga, Josip, et al. "On robustness and transferability of convolutional neural networks." arXiv:2007.08558. 2020.

Data requirements

- Define the features, their units and compute simple descriptive statistics
- Infer the data schema:
  - Data type
  - Presence (required/optional)
  - Valid range
- Code the tests for data validation
- Code how to deal with anomalies
- Monitor the data “drift”



People+AI research (PAIR) at Google
https://pair-code.github.io/facets/

Source: TensorFlow Data Validation | TFX

Datasheets for Datasets

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is accompanied by a label indicating the name of the person in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included in the dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The dataset comes with specified train/test splits such that none of the people in the training split are in the test split and vice versa. The data is split into two views, View 1 and View 2. View 1 consists of a training subset (pairsDevTrain.txt) with 1100 pairs of matched and 1100 pairs of mismatched images, and a test subset (pairsDevTest.txt) with 500 pairs of matched and mismatched images. Practitioners can train an algorithm on the training set and test on the test set, repeating as often as necessary. Final performance results should be reported on View 2 which consists of 10 subsets of the dataset. View 2 should only be used to test the performance of the final model. We recommend reporting performance on View 2 by using leave-one-out cross validation, performing 10 experiments. That is, in each experiment, 9 subsets should be used as a training set and the 10^{th} subset should be used for testing. At a minimum, we recommend reporting the **estimated mean accuracy,** $\hat{\mu}$ and the **standard error of the mean:** S_E for View 2.

$\hat{\mu}$ is given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \quad (1)$$

where p_i is the percentage of correct classifications on View 2 using subset i for testing. S_E is given as:

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \quad (2)$$

[1] All information in this datasheet is taken from one of five sources. Any errors that were introduced from these sources are our fault.
Original paper: http://www.cs.cornell.edu/people/pabo/movie-review-data/; LFW survey: http://vis-www.cs.umass.edu/lfw/lfw.pdf; Paper measuring LFW demographic characteristics : http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf; LFW website: http://vis-www.cs.umass.edu/lfw/.

Table 1 summarizes some dataset statistics and Figure 1 shows examples of images. Most images in the dataset are color, a few are black and white.

Property	Value
Database Release Year	2007
Number of Unique Subjects	5649
Number of total images	13,233
Number of individuals with 2 or more images	1680
Number of individuals with single images	4069
Image Size	250 by 250 pixels
Image format	JPEG
Average number of images per person	2.30

Table 1. A summary of dataset statistics extracted from the original paper: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%

Table 2. Demographic characteristics of the LFW dataset as measured by Han, Hu, and Anil K. Jain. *Age, gender and race estimation from unconstrained face images*. Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014).

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The names for each person in the dataset were determined by an operator by looking at the caption associated with the person's photograph.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley[3]. The images in this database were gathered from news articles on the web using software to crawl news articles.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The original Faces in the Wild dataset is a sample of pictures of people appearing in the news on the web. Labeled Faces in the Wild is thus also a sample of images of people found on the news on line. While the intention of the dataset is to have a wide range of demographic (e.g. age, race, ethnicity) and image (e.g. pose, illumination, lighting) characteristics, there are many groups that have few instances (e.g. only 1.57% of the dataset consists of individuals under 20 years old).

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Subsequent gender, age and race annotations listed in http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf were performed by crowd workers found through Amazon Mechanical Turk.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Unknown

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes. Each instance is an image of a person.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was crawled from public web sources.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Unknown

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No. All subjects in the dataset appeared in news sources so the images that we used along with the captions are already public.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

[3] Faces in the Wild: http://tamaraberg.com/faceDataset/

Uses

Has the dataset been used for any tasks already? If so, please provide a description.

Papers using this dataset and the specified evaluation protocol are listed in http://vis-www.cs.umass.edu/lfw/results.html

Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.

Papers using this dataset and the specified training/evaluation protocols are listed under "Methods" section of http://vis-www.cs.umass.edu/lfw/results.html

What (other) tasks could the dataset be used for?

The LFW dataset can be used for the face identification problem. Some researchers have developed protocols to use the images in the LFW dataset for face identification.[5]

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

There is minimal risk for harm: the data was already public.

Are there tasks for which the dataset should not be used? If so, please provide a description.

The dataset should not be used for tasks that are high stakes (e.g. law enforcement).

Any other comments?

Distribution

Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.

Yes. The dataset is publicly available.

How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?

The dataset can be downloaded from http://vis-www.cs.umass.edu/lfw/index.html#download. The images can be downloaded as a gzipped tar file.

When will the dataset be distributed?

The dataset was released in October, 2007.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The crawled data copyright belongs to the news papers that the data originally appeared in. There is no license, but there is a request to cite the corresponding paper if the dataset is used: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

There are no fees or restrictions.

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

Unknown

Any other comments?

[5] Unconstrained face recognition: Identifying a person of interest from a media collection: http://biometrics.cse.msu.edu/Publications/Face/BestRowdenetal.UnconstrainedFaceRecognition_TechReport_MSU-CSE-14-1.pdf

Maintenance

Who will be supporting/hosting/maintaining the dataset?

The dataset is hosted at the University of Massachusetts.

How can the owner/curator/manager of the dataset be contacted (e.g., email address)?

All questions and comments can be sent to Gary Huang: gbhuang@cs.umass.edu.

Is there an erratum? If so, please provide a link or other access point.

All changes to the dataset will be announced through the LFW mailing list. Those who would like to sign up should send an email to lfw-subscribe@cs.umass.edu. Errata are listed under the "Errata" section of http://vis-www.cs.umass.edu/lfw/index.html

Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

All changes to the dataset will be announced through the LFW mailing list.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

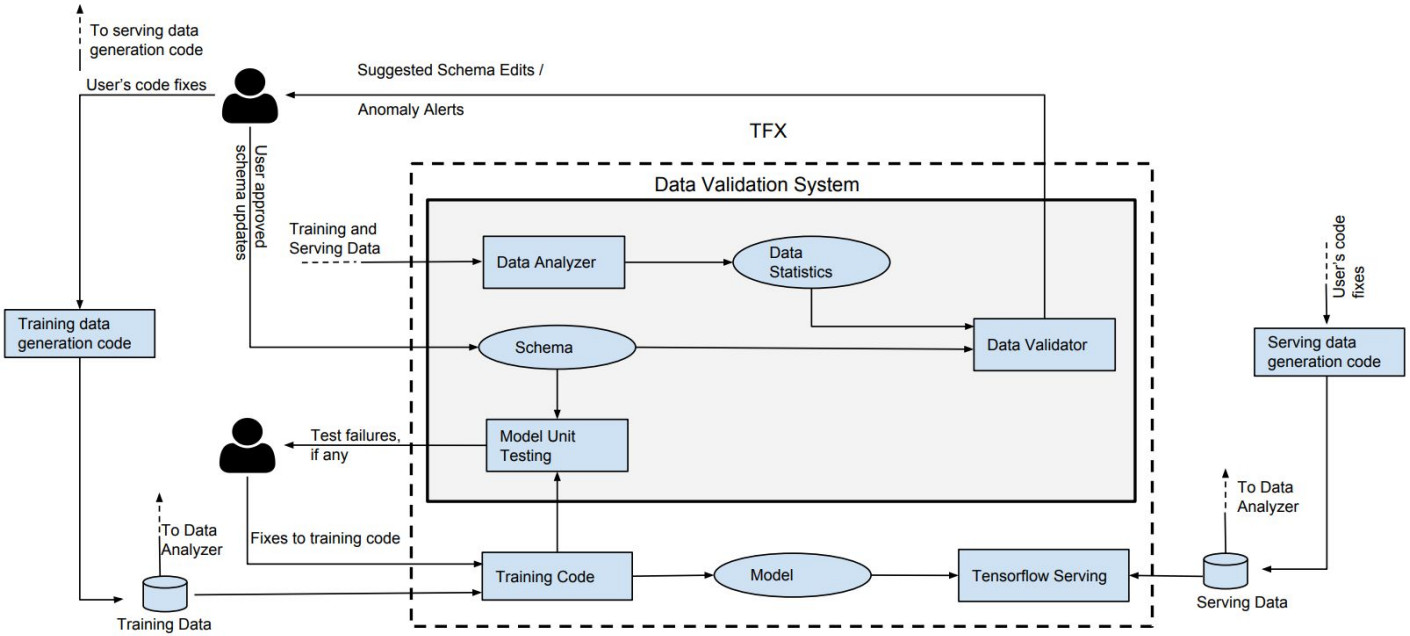
They will continue to be supported with all information on http://vis-www.cs.umass.edu/lfw/index.html unless otherwise communicated on the LFW mailing list.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

Gebru, Timnit, et al. "Datasheets for datasets." arXiv preprint arXiv:1803.09010 (2018).

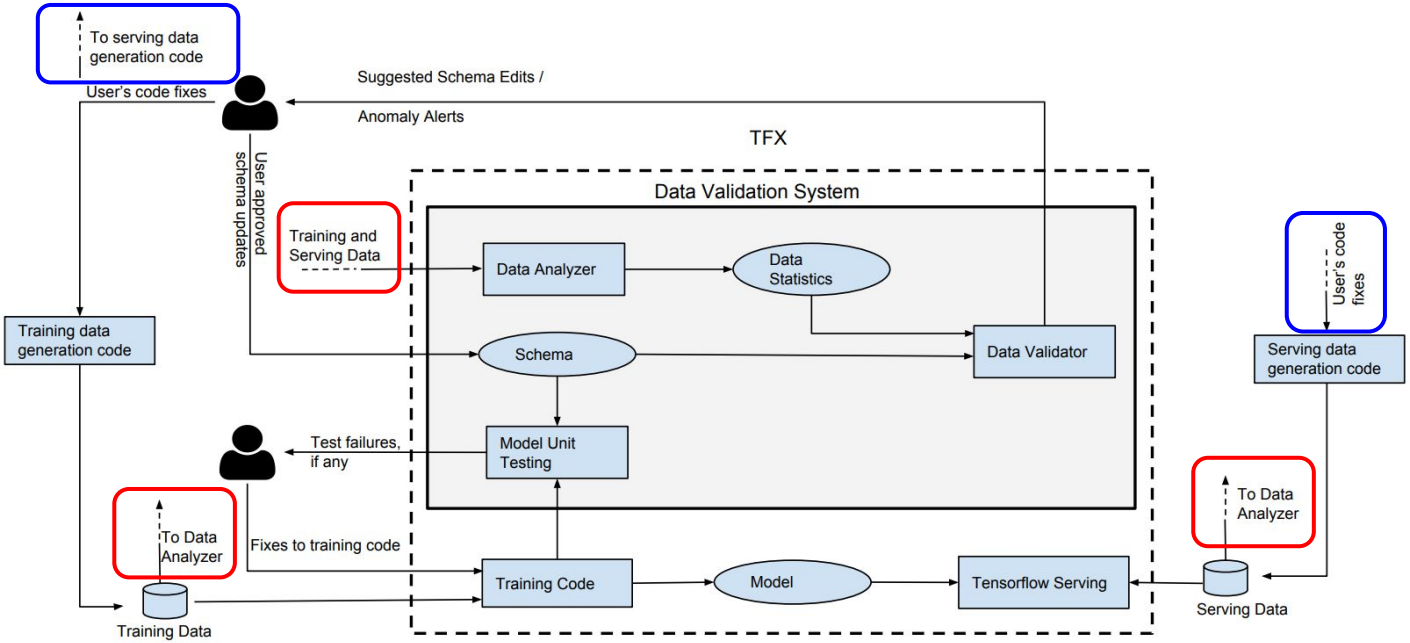
Mila

Data requirements: automating the validation



Polyzotis, Neoklis, et al. "Data validation for machine learning." Proceedings of Machine Learning and Systems. 2019: 334-347.

Data requirements: automating the validation



Source: Polyzotis, Neoklis, et al. "Data validation for machine learning." Proceedings of Machine Learning and Systems (2019): 334-347.

How to build trustworthy AI?

"AI developers should be explicit about the contracts that their models maintain."

European Guidelines for Trustworthy AI Models		Documentations	Explanatory Methods/Analyses
Key Requirements	*Factors*		
Human agency and oversight	Foster fundamental human rights	Fairness checklists	See "Diversity, non-discrimination, fairness"
	Support users' agency	All	User-centered explanations [59]
	Enable human oversight	-	Explanations in recommender systems [41]
Technical robustness and safety	Resilience to attack and security	Factsheets (security)	Adversarial attacks and defenses [20]
	Fallback plan and general safety	-	-
	A high level of accuracy	Model cards (metrics)	-
	Reliability	Factsheets (concept drift)	Contrast sets [17], behavioral testing [58]
	Reproducibility	Reproducibility checklists	"Show your work" [14]
Privacy and data governance	Ensure privacy and data protection	Datasheets/statements	Removal of protected attributes [57]
	Ensure quality and integrity of data	Datasheets/statements	Detecting data artifacts [23]
	Establish data access protocols	Datasheets/statements	-
Transparency	High-standard documentation	All	-
	Technical explainability	Factsheets (explainability)	Saliency maps [61], self-attention patterns [40], influence functions [38], probing [16]
	Adaptable user-centered explainability	Factsheets (explainability)	Counterfactual [21], contrastive [51], natural language [27], by-example [38], and concept-level [19] explanations
	Make AI systems identifiable as non-human	-	-
Diversity, non-discrimination, fairness	Avoid unfair bias	Fairness checklists	Debiasing using data manipulation [66]
	Encourage accessibility and universal design	-	-
	Solicit regular feedback from stakeholders	Fairness checklists	-
Societal and environmental well-being	Encourage sustainable and eco-friendly AI	Reproducibility checklists	Analayzing individual neurons [10]
	Assess the impact on individuals	Fairness checklists	Bias exposure [65]
	Assess the impact on society and democracy	Fairness checklists	Explanations designed for applications such as fact checking [3] or fake news detection [47]
Accountability	Auditability of algorithms/data/design	Factsheets (lineage)	-
	Minimize and report negative impacts	Fairness checklists	-
	Acknowledge and evaluate trade-offs	-	Reporting the robustness-accuracy [1] trade-off or simplicity-equity trade-off [37]
	Ensure redress	Fairness checklists	-

Source: Jacovi, Alon, et al. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI." arXiv preprint arXiv:2010.07487 (2020).

How to build trustworthy AI?

"AI developers should be explicit about the contracts that their models maintain."

European Guidelines for Trustworthy AI Models		Documentations	Explanatory Methods/Analyses
Key Requirements	*Factors*		
Human agency and oversight	Foster fundamental human rights	Fairness checklists	See "Diversity, non-discrimination, fairness"
	Support users' agency	All	User-centered explanations [59]
	Enable human oversight	-	Explanations in recommender systems [41]
Technical robustness and safety	Resilience to attack and security	Factsheets (security)	Adversarial attacks and defenses [20]
	Fallback plan and general safety	-	-
	A high level of accuracy	Model cards (metrics)	-
	Reliability	Factsheets (concept drift)	Contrast sets [17], behavioral testing [58]
	Reproducibility	Reproducibility checklists	"Show your work" [14]
Privacy and data governance	Ensure privacy and data protection	Datasheets/statements	Removal of protected attributes [57]
	Ensure quality and integrity of data	Datasheets/statements	Detecting data artifacts [23]
	Establish data access protocols	Datasheets/statements	-
Transparency	High-standard documentation	All	-
	Technical explainability	Factsheets (explainability)	Saliency maps [61], self-attention patterns [40], influence functions [38], probing [16]
	Adaptable user-centered explainability	Factsheets (explainability)	Counterfactual [21], contrastive [51], natural language [27], by-example [38], and concept-level [19] explanations
	Make AI systems identifiable as non-human	-	-
Diversity, non-discrimination, fairness	Avoid unfair bias	Fairness checklists	Debiasing using data manipulation [66]
	Encourage accessibility and universal design	-	-
	Solicit regular feedback from stakeholders	Fairness checklists	-
Societal and environmental well-being	Encourage sustainable and eco-friendly AI	Reproducibility checklists	Analayzing individual neurons [10]
	Assess the impact on individuals	Fairness checklists	Bias exposure [65]
	Assess the impact on society and democracy	Fairness checklists	Explanations designed for applications such as fact checking [3] or fake news detection [47]
Accountability	Auditability of algorithms/data/design	Factsheets (lineage)	-
	Minimize and report negative impacts	Fairness checklists	-
	Acknowledge and evaluate trade-offs	-	Reporting the robustness-accuracy [1] trade-off or simplicity-equity trade-off [37]
	Ensure redress	Fairness checklists	-

Jacovi, Alon, et al. "Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI." arXiv preprint arXiv:2010.07487 (2020).

SOTA NLP models are brittle and spurious

- The metrics are flawed (BLEU, ROUGE, ...)
- The datasets are too limited
  - Heuristics can work pretty well [1]
  - Not ecologically valid research [2]
    - Synthetic language
    - artificial tasks
    - Not working with prospective users
    - use of script and/or priming
    - singe-turn interfaces
    - Automatic evaluation

“We find that four existing NLI models perform very poorly on HANS, suggesting that their high accuracies on NLI test sets may be due to the exploitation of invalid heuristics rather than deeper understanding of language” [1]

[1] Tom McCoy, Ellie Pavlick, Tal Linzen: Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ACL (2019)
[2] de Vries, Harm, Dzmitry Bahdanau, and Christopher Manning. "Towards ecologically valid research on language user interfaces." arXiv:2007.14435 (2020)

Robustification

	ImageNet-200 (%)	ImageNet-R (%)	Gap
ResNet-50	7.9	63.9	56.0
+ ImageNet-21K *Pretraining* (10× labeled data)	7.0	62.8	55.8
+ CBAM (*Self-Attention*)	7.0	63.2	56.2
+ ℓ_∞ Adversarial Training	25.1	68.6	43.5
+ Speckle Noise	8.1	62.1	54.0
+ Style Transfer Augmentation	8.9	58.5	49.6
+ AugMix	7.1	58.9	51.8
+ DeepAugment	7.5	57.8	50.3
+ DeepAugment + AugMix	8.0	53.2	45.2
ResNet-152 (*Larger Models*)	6.8	58.7	51.9

Table 1: ImageNet-200 and ImageNet-R top-1 error rates. ImageNet-200 uses the same 200 classes as ImageNet-R. DeepAugment+AugMix improves over the baseline by over 10 percentage points. ImageNet-21K Pretraining tests *Pretraining* and CBAM tests *Self-Attention*. Style Transfer, AugMix, and DeepAugment test *Diverse Data Augmentation* in contrast to simpler noise augmentations such as ℓ_∞ Adversarial Noise and Speckle Noise. While there remains much room for improvement, results indicate that progress on ImageNet-R is tractable.

Source: Hendrycks, Dan, et al. "The many faces of robustness: A critical analysis of out-of-distribution generalization." arXiv preprint arXiv:2006.16241 (2020).

Real-world problems

4.1. Key Findings on Evaluated Classifiers

- All classifiers perform better on male faces than female faces (8.1% – 20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8% – 19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8% – 34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

RETAIL OCTOBER 10, 2018 / 7:04 PM / UPDATED 2 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin 8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's AMZN.O machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Out-of-distribution

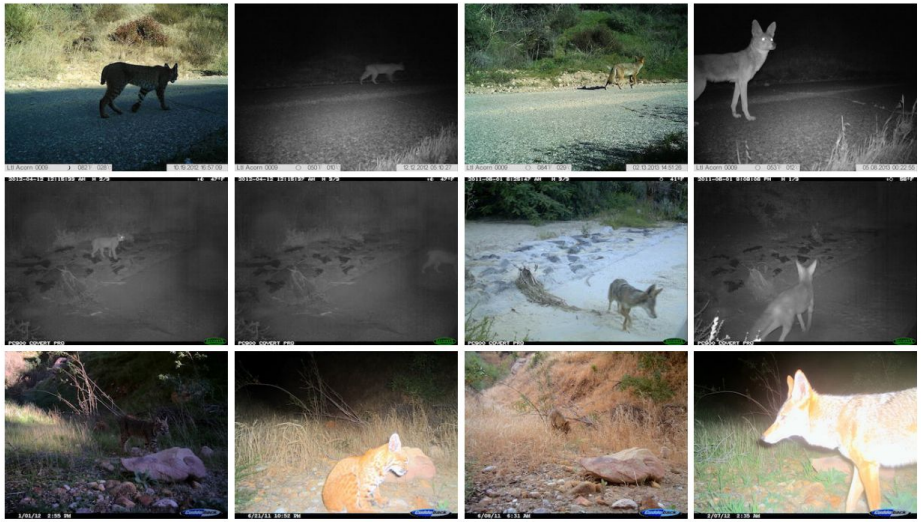
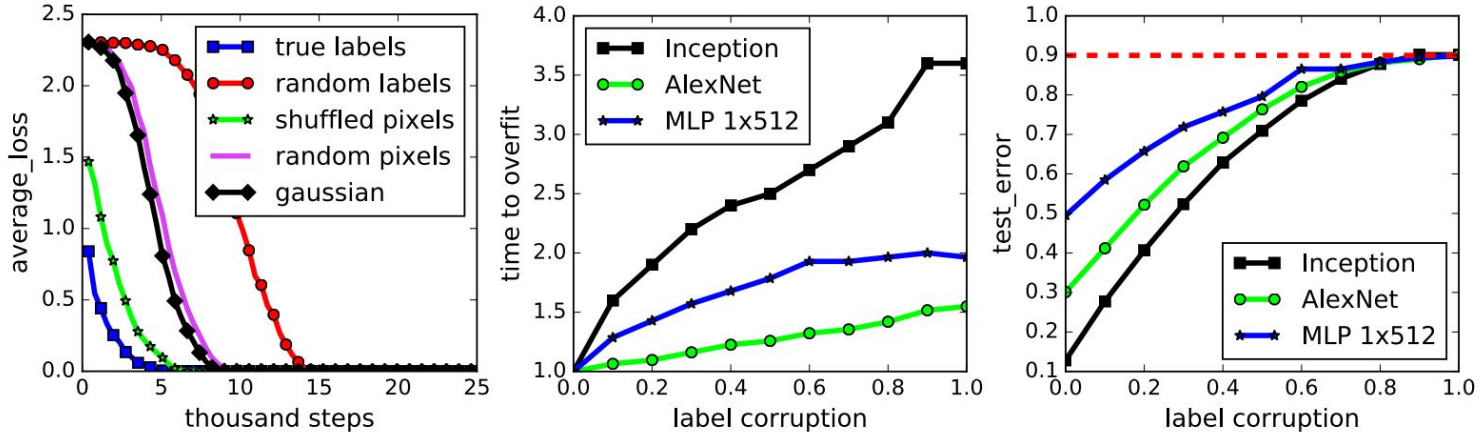


Fig. 2. Camera trap images from three different locations. Each row is a different location and a different camera type. The first two cameras use IR, while the third row used white flash. The first two columns are bobcats, the next two columns are coyotes.

Source: Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita." Proceedings of the European Conference on Computer Vision (ECCV). 2018.

Lazy associative machines



(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

Source: Zhang, Chiyuan, et al. "Understanding deep learning requires rethinking generalization." arXiv preprint arXiv:1611.03530 (2016).

Calibration

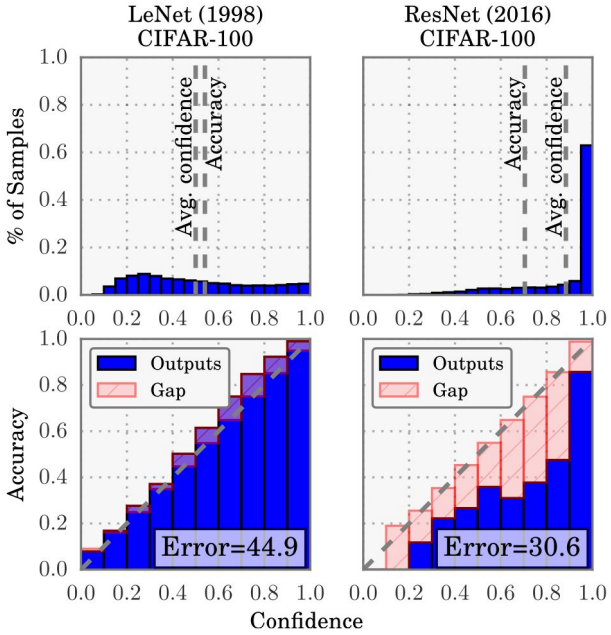


Figure 1. Confidence histograms (top) and reliability diagrams (bottom) for a 5-layer LeNet (left) and a 110-layer ResNet (right) on CIFAR-100. Refer to the text below for detailed illustration.

Source: Chuan Guo, Geoff Pleiss, Yu Sun, Kilian Q. Weinberger: On Calibration of Modern Neural Networks. ICML 2017: 1321-1330