Advanced Deep Learning: Explainability

Anders Søgaard

coAStaL

Course outline

Goal 1: Quick tour of recent developments in deep learning

Goal 2: Inspiration for thesis/research projects

Week	Lecturer	Subject	Literature	Assignment
1	Stefan	Introduction to Neural Networks.	d2l 2.1-2.5, 2.7, 11.5.1, **slides**	
2	Stefan	CNNs; FCNs; U-Nets. Data augmentation; invariance; regularization e.g. dropout	d2l 6, 7, 13.9-13.11, **slides**	Assignment 1 (May 10)
3	Anders/Phillip	May 9 (A): RNNs May 11 (P): Transformers	d2l 8 Transformers: d2l 10.5-10.7 + Vaswani et al. (2017)	Assignment 2 (May 20)
4	Phillip/Anders	May 16 (P): Representation and Adversarial Learning May 18 (A): A Learning Framework + Self-supervised Learning + Contrastive Learning	Autoencoders: blog post GANs: Goodfellow (2016) Self-supervised learning: blog post Contrastive learning: Dor et al. (2018) Adversarial examples: Goodfellow et al. (2015)	
5	Anders	May 23: General Properties, e.g., Scaling Laws, Lottery Tickets, Bottleneck Phenomena May 25: Applications of Representation, Adversarial and Contrastive Learning	GANs: Lample et al. (2018) Autoencoders: Chandar et al. (2011) Contrastive learning: Yu et al. (2018) DynaBench: Talk by Douwe Kiela (Facebook, now HuggingFace) Scaling laws: Kaplan et al. (2020)	Assignment 3 *[MC on Representation Learning/1p Report on Lottery Ticket extraction]* (June 3)
6	Anders	May 30: Interpretability, Transparency, and Trustworthiness & Deep Learning for Scientific Discovery June 1: Interpretability (Feature Attribution), including Guest Lecture by Stephanie Brandl	DL for Scientific Discovery: Sullivan (2022) Interpretability/Background: Søgaard (2022)	
7	Anders	June 6: *Off (no teaching)* June 8: Interpretability (Training Data Influence)	Literature: Feng and Boyd-Graber (2018); Jiang and Senge (2021)	
8	Anders	June 13-15: Best Practices	Literature: Dodge et al. (2019) and Raji et al. (2021)	Assignment 4 *[MC on Interpretability; 1p Report on Best Practices]* (June 21)

Course outline

Goal 1: Quick tour of recent developments in deep learning

Goal 2: Inspiration for thesis/research projects

Architectures

Framework

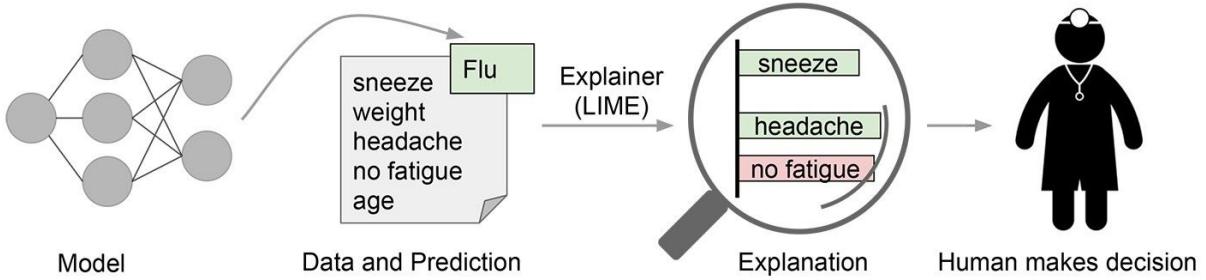
Fairness / Explainable AI

Methodology

Week	Lecturer	Subject	Literature	Assignment
1	Stefan	Introduction to Neural Networks.	d2l 2.1-2.5, 2.7, 11.5.1, **slides**	
2	Stefan	CNNs; FCNs; U-Nets. Data augmentation; invariance; regularization e.g. dropout	d2l 6, 7, 13.9-13.11, **slides**	Assignment 1 (May 10)
3	Anders/Phillip	May 9 (A): RNNs May 11 (P): Transformers	d2l 8 Transformers: d2l 10.5-10.7 + Vaswani et al. (2017)	Assignment 2 (May 20)
		May 16 (P): Representation and Adversarial Learning	Autoencoders: blog post GANs: Goodfellow (2016)	
4	Phillip/Anders	May 18 (A): A Learning Framework + Self-supervised Learning + Contrastive Learning	Self-supervised learning: blog post Contrastive learning: Dor et al. (2018) Adversarial examples: Goodfellow et al. (2015)	
5	Anders	May 23: General Properties, e.g., Scaling Laws, Lottery Tickets, Bottleneck Phenomena May 25: Applications of Representation, Adversarial and Contrastive Learning	GANs: Lample et al. (2018) Autoencoders: Chandar et al. (2011) Contrastive learning: Yu et al. (2018) DynaBench: Talk by Douwe Kiela (Facebook, now HuggingFace) Scaling laws: Kaplan et al. (2020)	Assignment 3 *[MC on Representation Learning/1p Report on Lottery Ticket extraction]* (June 3)
6	Anders	May 30: Interpretability, Transparency, and Trustworthiness & Deep Learning for Scientific Discovery June 1: Interpretability (Feature Attribution), including Guest Lecture by Stephanie Brandl	DL for Scientific Discovery: Sullivan (2022) Interpretability/Background: Søgaard (2022)	
7	Anders	June 6: *Off (no teaching)* June 8: Interpretability (Training Data Influence)	Literature: Feng and Boyd-Graber (2018); Jiang and Senge (2021)	
8	Anders	June 13-15: Best Practices	Literature: Dodge et al. (2019) and Raji et al. (2021)	Assignment 4 *[MC on Interpretability; 1p Report on Best Practices]* (June 21)

Today

Last time: Feature attribution methods and how (not) to evaluate them.



Today

a) Training data influence (local)
b) How (not) to evaluate influence-based approaches
c) Human evaluation
d) Stephanie's talk
e) Probing classifiers (global)
f) Sullivan (2022)'s criticism of XAI methods (that they are irrelevant to scientific discovery)

Training data influence

Grad-Cos

Baseline method: Simply returns the cosine distance of the gradients of *z* and *z'*.

Note: Common alternative is Grad-Dot.

Figure 6: Top 5 influential points for the test point: 1479 (CIFAR-10). The model is a ResNet-18 trained with a weight-decay regularization; Only 3 out of the 5 points are semantically similar to the test-point with class "Bird".

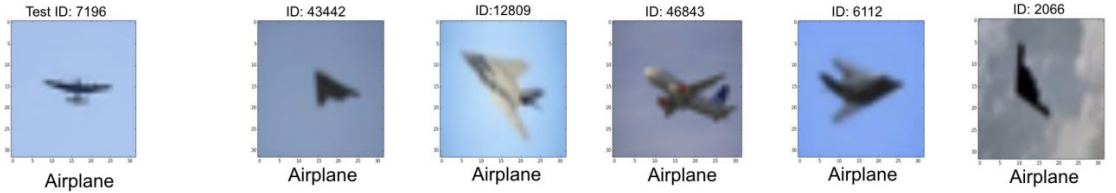


Figure 7: Top 5 influential points for the test point: 7196 (CIFAR-10). The model is a ResNet-18 trained with a weight-decay regularization; All the 5 training points are semantically similar to the test-point from the class "Airplane".

Influence functions

An old technique for quantifying how the model parameters change as we upweight a training point by an infinitesimal amount.

Problems: Expensive and only works for convex models.

Solution: Approximations.

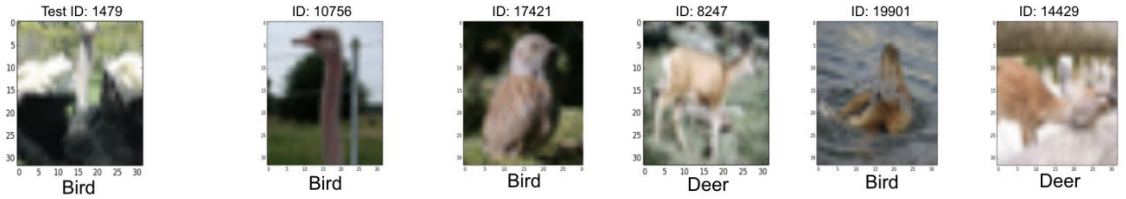


Figure 6: Top 5 influential points for the test point: 1479 (CIFAR-10). The model is a ResNet-18 trained with a weight-decay regularization; Only 3 out of the 5 points are semantically similar to the test-point with class "Bird".

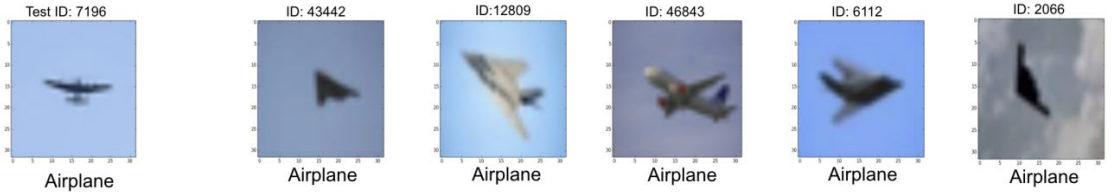
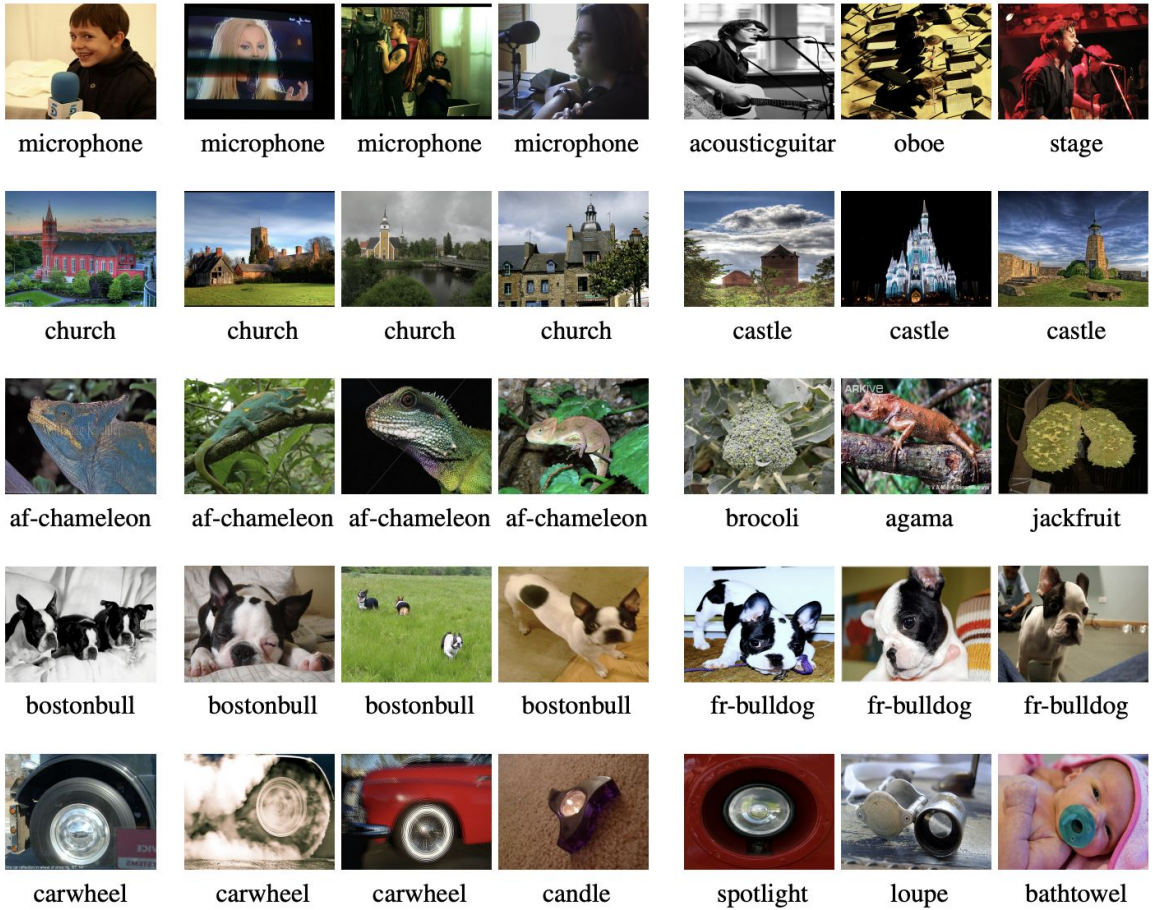


Figure 7: Top 5 influential points for the test point: 7196 (CIFAR-10). The model is a ResNet-18 trained with a weight-decay regularization; All the 5 training points are semantically similar to the test-point from the class "Airplane".

TraceInCP

Store check-points. Make influence of a training data point (z) on a test data point (z'):

$$\texttt{TracInCP}(z, z') = \sum_{i=1}^{k} \eta_i \nabla \ell(w_{t_i}, z) \cdot \nabla \ell(w_{t_i}, z')$$



Evaluating Training Data Influence

Leave-one-out influence

- Train a model for all *n*-1 subsets of your *n*-sized training data
- The influence of *z* is the difference in output on *z'* between the model trained on all data - and the model trained on all data but {*z*}
- Often considered gold standard

Exercise: What's the problem with using leave-one-out influence as a gold standard?

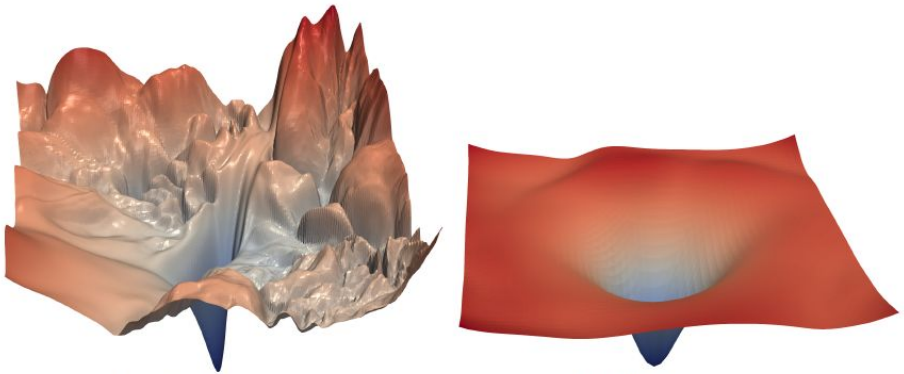
Data dependence

The way influence functions are used sometimes suggests influence is a stable relation between training and test examples. One example is Kocijan and Bowman (2020), who use influence scores to resample data for transfer learning. They obtain negative results, which we believe are easily explained by the observation in the above that influence scores are unstable across initializations and reorderings. That is, there is no guarantee that the data with high influence scores is useful in the context of a new model, trained on a subset of the original data.

In sum, influence depends on the data distribution.

Initialization

Two models may be identified from different initializations.



(a) without skip connections (b) with skip connections

A training data point may have had influence in leading to the optimum from one initialization, but not from another.

Heuristics

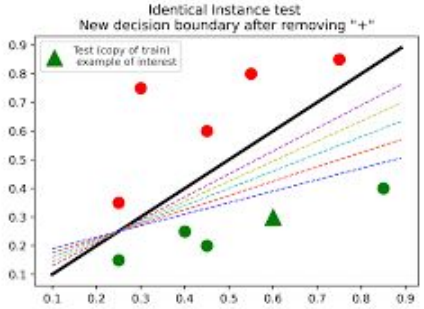
- The amount of the training data points that have themselves as most influential
- The amount of the test data points in class agreement with their most influential

Exercise: What's the problem with these heuristics? Think of a counter-example to both.

Heuristics

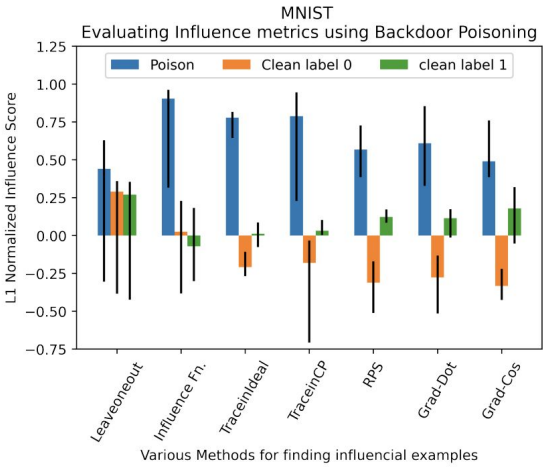
- The amount of the training data points that have themselves as most influential
- The amount of the test data points in class agreement with their most influential

Exercise: What's the problem with these heuristics? Think of a counter-example to both.

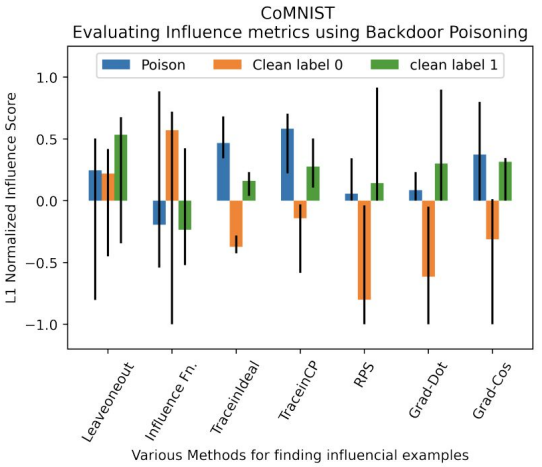


Backdoor poisoning attacks

We add a trigger or poison signature to a small portion of training examples from class A and relabel them as class B; then we train the model using this poisoned training data. If the model predicts poisoned test samples as B (and unpoisoned correspondents as A), it can only be because of the poisoned training data. From this, we can derive a metric for influence-based methods.



(a) MNIST



(b) CoMNIST

Human Evaluation

QuizBowl & Jeopardy

Can a computer help a human play QuizBowl or Jeopardy? - and if so, would it help if it supplied rationales for its advice?

Quick story

- Nguyen (2018) first to do forward prediction evaluation in NLP.

disaster films have a tendency to be very formulated and very **cliched** . to see a disaster film with actual **originality** , or at least a **decent** plot twist , would definitely be a welcome surprise . **unfortunately** , **folks** , it's not **likely** . dante's peak is **cliched** , and at times corny , but also pretty **decent** . to be honest , i **wasn**'t very interested in seeing this film , and word of mouth , as well as several reviews , didn't make it sound **promising** . so i was pleasantly surprised that to find that this movie **wasn**'t bad at all . it's pretty **run** of the mill , but it's not something i would say is **merely** " **ok** " to watch . in case you don't know , dante's peak is about a volcano and the city which lives in it's shadow , dante's peak (who would've guessed , eh ?) . pierce brosnan plays the volcanologist sent to study the volcano and , perhaps more by hunch than actual scientific proof , is determined that the volcano will be arupting in the very near future . due to the **lack** of more substantial **evidence** , nobody warns the small town , and when they finally do , it's in the **middle** of the town meeting that the volcano finally **blows** . brosnan , all around good guy , will , of course , **save** the day . . . or at least the mayor of dante's peak (linda hamilton) . naturally the two will become infatuated with one another . (if you think i just ruined a plot development , you haven't seen very many movies !) there's also the virtually neccessary kids and **pet dog** to tug at your heart strings . and of course , the kids or the **dog** (at least one or the other) will do something heroic . . . but hey , i don't want to ruin all the surprises ! if there was a part of you that was hesitating seeing dante's peak **merely** because it was rumored to be a **waste** of time , i urge you to watch it and decide for yourself . it's not brain food , but it succeeds at what it's **meant** to be . . . an enjoyable , suspenseful movie about the fury mother nature can unleash !

Choose the system output: (required)

- ○ Positive
- ○ Negative

I am confident in my answer: (required)

	1	2	3	4	5	
Strongly disagree	○	○	○	○	○	Strongly agree

Quick story

- Nguyen (2018) first to do forward prediction evaluation in NLP.
- *Exercise:* What's the problem with her design?

Quick story

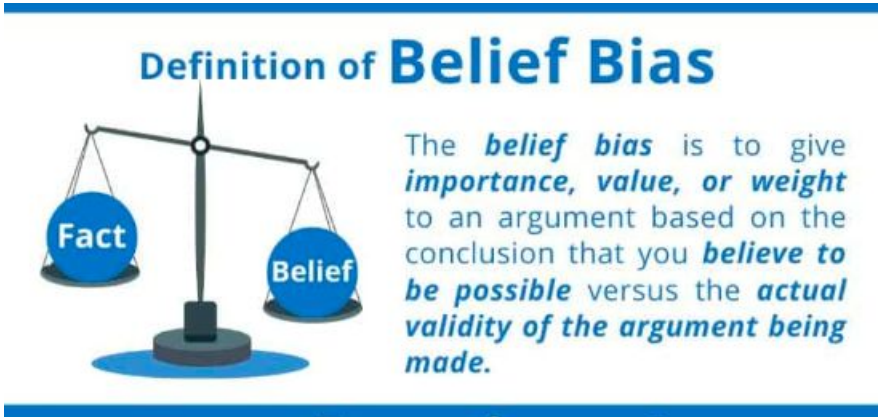
- Nguyen (2018) first to do forward prediction evaluation in NLP.
- Given participants have little experience with the model, all they can do is
  - a) say whether they find the marked passages (rationales) positive or negative (on their own account)
  - b) read the rest of the review (which is irrelevant for evaluating the rationales) for additional support.

Quick story

- Nguyen (2018) first to do forward prediction evaluation in NLP.
- This design is therefore extremely prone to human biases.
- Hase and Bansal (2020) improve on this design by introducing a training phase.
- Gonzalez and Søgaard (2021) improve the design further by also masking the task.

Belief bias

This is because **humans rely on belief bias and predict the gold standard answer more often than the model prediction**.



Model selection task

Gonzalez et al. (2021) introduce a second evaluation protocol, in addition to human forward prediction, namely model selection.