Vision-Language Pretraining: Current Trends and the Future

Part 3: Beyond statistical learning

Damien Teney

ACL Tutorial, 22 May 2022

Let's take a step back...

We use machine learning mostly to build predictive models that imitate real-world systems.

Statistical models exploit any predictive correlation to make the best predictions, regardless of its relevance to the task.

- **Image recognition**: imitate a human labeller.

  Predictive correlation: blue background / bird.

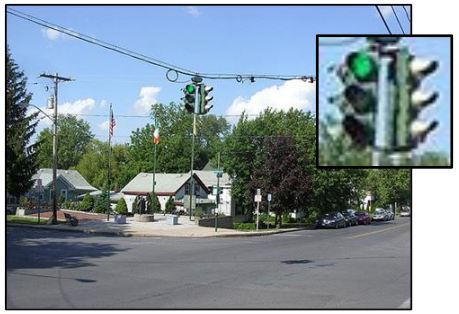
- **Visual question answering** (VQA): imitate a human answering questions.

  Predictions can be correct for the wrong reasons.

  E.g. because the model relies on a (biased!) prior distribution of answers.

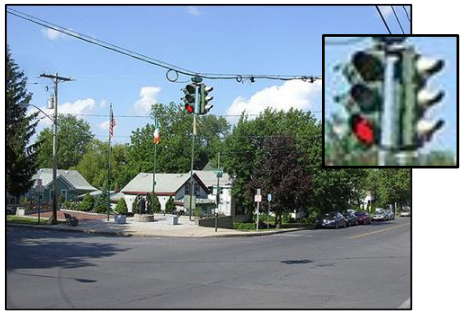
How many ...	2 (or 3)
Is/Are ...	"Yes" (~80%) vs "no"
What sport ...	Tennis
What animal ...	Dog
What is the color of ...	Red, blue

What color is illuminated on the traffic light ?



Predicted answer: *Green*. ✓

Testing with an edited image:



Predicted answer: *Green*. ✗

Statistical learning has limitations.

Predictions are only reliable within the training distribution.

Training data (biased)

Challenging if the model relies on grass in the background.

Test data (out-of-distribution)

The features used by a model are not necessarily those the real system (the human labeller) relies on.

E.g. the background is not causal to the predicted variable (label).

⟺ *Intervening on an image by changing its background would not cause one to label it differently.*

A cow.

Intervention →

Still a cow.

More limitations: statistical models only answer predictive questions.

- Example: a model **predicting whether a machine in a factory is about to fail** based on the noise it makes.

  The model can't tell how to reduce the rate of failures.

  The noise is not what causes the machine to fail (noise/failure is only a correlation). Obvious by common sense.

- Example: an NLP model **predicting the popularity (future number of clicks)** of news articles.

  Interpretability methods may show that the model relies on headline length.

  But it doesn't mean we can alter popularity by changing headline lengths (only a correlation in tr. data). Not as obvious !

Even if we care only about predictive questions, we also want:

- Adversarial robustness *= Worst-case OOD generalization.*
- Compositional and cross-task generalization *= Repurposing bits of learned knowledge.*
- Accurate predictions in conditions not seen in training

All these settings violate the assumption in statistical learning of i.i.d. training/test data.

Causality can help ML to overcome these limitations.

- Notations and principles to describe the data-generating mechanisms (more fundamental than observed correlations).
  Whereas the language of statistics (e.g. conditional probabilities) can only describe observational properties.

- Data-generating mechanisms are more fundamental than observed correlations.
  We can derive correlations from causal structure, but not the reverse.

- "A causes B" ≡ Ⓐ→Ⓑ ≡ Setting A to a specific value can affect the distribution of B.
  $P(B|do(A)) \neq P(B|A)$ A new 'do' operator represents interventions.

- Causal learning = learning the data-generating mechanisms that define a task (not just a specific dataset).

This talk:
- **Causal language/principles to help you navigate the literature on your own.**
- **Example applications: evaluating the robustness of V&L models.**
- **Example applications: training better models.**

Generalization ≠ generalization

In-domain (ID) generalization

- Classical use of the term.
- Easier with more training data.
- Inductive biases are indispensable, but some seem universally useful.

  E.g. simple regularizers that favour simplicity/smoothness such as weight decay.

Out-of-distribution (OOD) generalization

- Some correlations from the training data may be absent or misleading at test time.
- More (of the same) data is not sufficient.
- Need task-specific information.

  E.g. should we rely on texture/shape/color for object recognition ?

The disciplines of causality

Classical methods (rely on predefined variables & explicit causal graphs)

- Causal inference: answering causal queries from observational data and a known causal graph (human-provided).

  How much would people click on ads if we double the font size ? *(provided a causal graph, and data from controlled and/or non-controlled experiments)*

  Would this specific person be in better health, had she been administered treatment X ?

- Causal discovery: using data to refine a partially-known causal graph.

  Is gene X responsible for health condition Y ?

Emerging area: extending ML with causal principles (high-dimensional data & causal relationships not modelled explicitly)

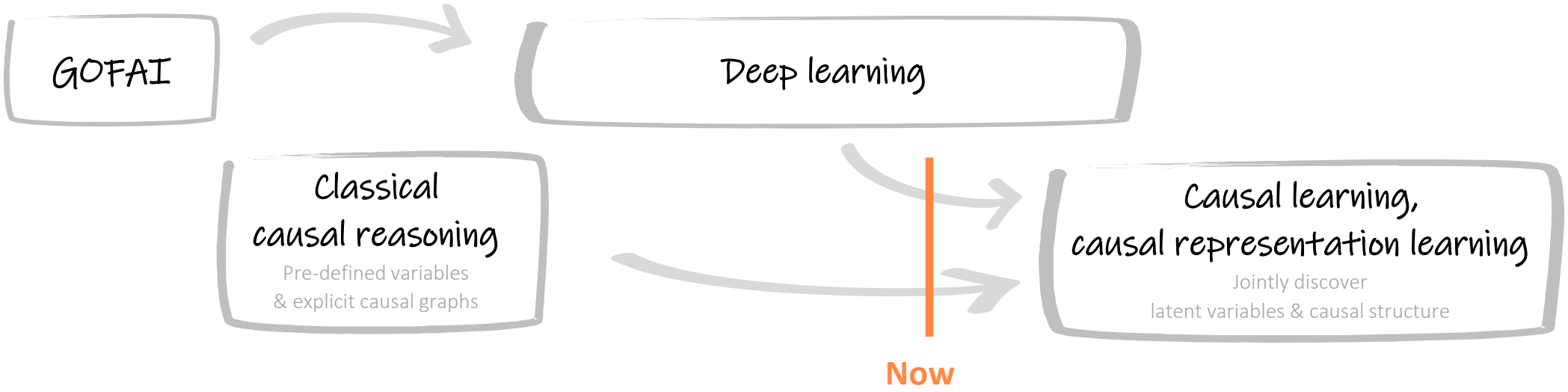
- Causal representation learning: learning embeddings of raw data, disentangling its generative factors (causal parents).

  Equivalent to: disentanglement, independent component analysis (ICA).

- Causal learning: learning predictive models that rely on causal (not spurious) features.

  Enable better transfer to unseen conditions, across datasets, across tasks.

  Also aims at (implicitly) identifying generative factors.

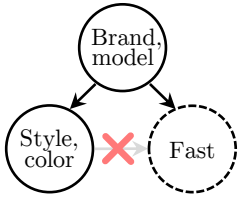

GOFAI
Deep learning
Classical
causal reasoning
Pre-defined variables
& explicit causal graphs
Causal learning,
causal representation learning
Jointly discover
latent variables & causal structure
Now

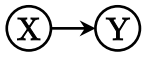
Statistical vs. causal model

"Is this a fast car ?"

Training images with labels 'Fast' ∈ {0,1}

Someone's mental (causal) model:



X causes Y 

⇔ Intervening on X affects Y.

⇔ $\mathrm{P}(\mathrm{Y}|\mathrm{do}(\mathrm{X}{=}x)) \neq \mathrm{P}(\mathrm{Y})$.

Conditioned on observing the color in the training distribution.

$$\mathrm{P}(Fast \mid Color)$$

$$\neq$$

$$\mathrm{P}(Fast \mid \mathrm{do}(Color))$$

Conditioned on an intervention.

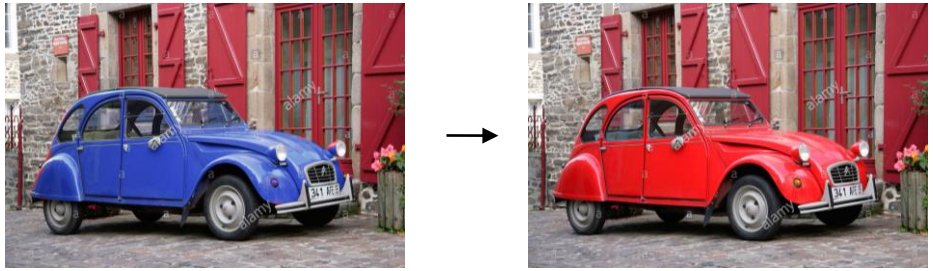
› **Statistical learning is about correlations:** red = fast.

Reliable only if the training/test data are from similar distributions.

› **Causal learning is about mechanisms.**

Enables predictions in conditions unobserved during training (OOD).

What happens to a re-painted car ?



Faster ? No !

› Learning such a model = mimicking the causal model of a real-world process

(e.g. a human answerer)

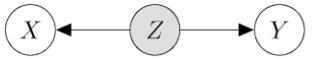
› **More data cannot distinguish spurious correlations from causal ones.**

(more red Ferraris)

Causal learning is difficult because we usually have only observational data. (passively collected)

- Would be easy if we could freely interact with the real world (just act and observe the effects!).
- **More observational data** does not help.
  Biased/long-tailed distributions remain biased/long-tailed, even with lots of samples !
- A typical dataset provides **i.i.d. samples of a joint distribution** e.g. over images/labels.
  Without assumptions or domain knowledge, spurious/robust correlations are indistinguishable.
- A joint distribution is usually compatible with **multiple causal structures**.
  A correlation between X and Y could arise from 

or 

or 

(hidden common cause).
- The causal structure is more fundamental than statistical correlations.
  We can derive correlations from the causal structure, but generally not the opposite.
  We need background knowledge, additional assumptions, or interventional/heterogeneous data.

Back to VQA…

Real world

Learned model

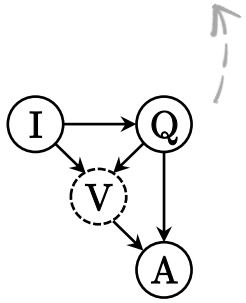
We want to mirror the causal mechanisms.

Real world

In reality, way too complex to describe as a graph

- The data-generating process that we seek to emulate.

  In VQA, **a human annotator** who takes an Image and Question, finds relevant Visual features then gives an Answer.



- A standard dataset only provides i.i.d. samples from the joint distribution.

$$\mathrm{P(I,Q,V,A)} = \overbrace{\mathbf{P(A|V,Q)}\ \mathbf{P(V|I,Q)}\ \mathrm{P(Q|I)}\ \mathrm{P(I)}}^{\text{Causal factorization}}$$

The joint distribution does not carry causal information.
Multiple causal structures can produce a same statistical signature.

Learned model

- The inference also has a causal structure (inputs → … → predictions).

- Only 2 options to obtain knowledge of the data-generating process.

  - **Human-provided, task-specific knowledge.** (assumptions, inductive biases)

    Examples: special architectures, hand-designed data augmentations, interaction w/ simulated human-designed environments.

  - **Other data carrying causal information.** (i.i.d. samples are not enough !)

    Examples: multiple training datasets/environments, counterfactual examples, non-stationary time series.

How to evaluate models ?

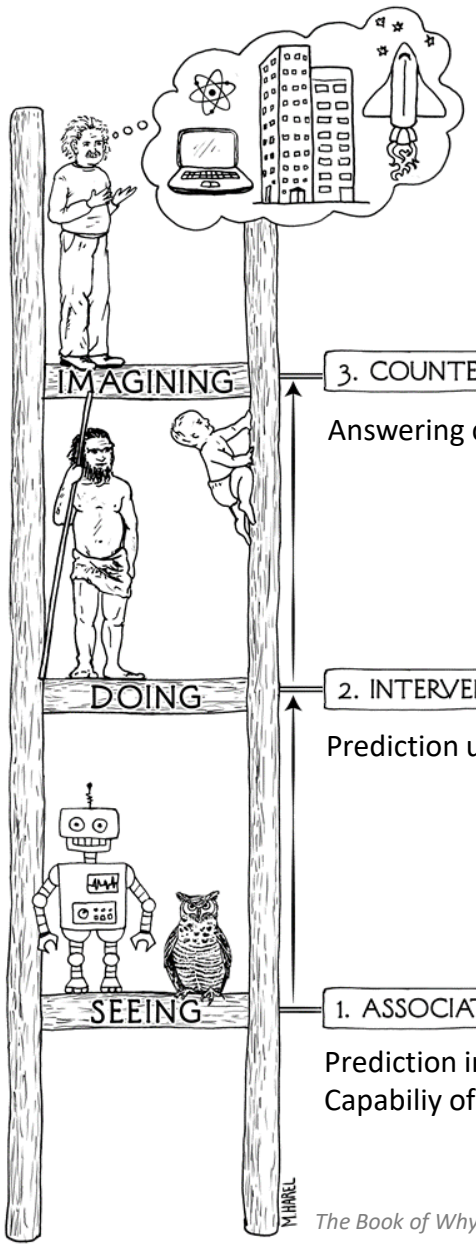
(from a causal perspective)

Does a VQA model

- reliably answer questions about novel/unusual (OOD) scenes ?
- rely on the same features as a human ?
- implement the same causal mechanisms as the real world ?

Note: Interpretability methods serve to compare (qualitatively) the causal structure of a model with our (mental) causal model of the world (i.e. what predictive features should be used). Here, we rather aim to do this quantitatively and with data.

Causal hierarchy: 3 types of queries to a model. ⇔ Evaluation settings used in machine learning.



Increasing difficulty →

3. COUNTERFACTUALS

Answering counterfactual questions.

2. INTERVENTION

Prediction under distrib. shift.

1. ASSOCIATION

Prediction in i.i.d. settings.
Capabiliy of statistical models.

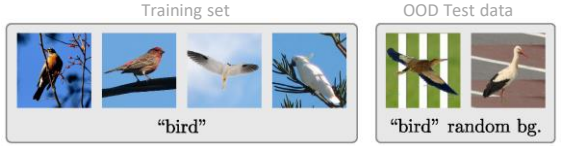
The Book of Why, Pearl & Mackenzie, 2019.

› Pairs of counterfactual test examples. (a.k.a. contrast sets)
Interventions at instance level. Probe models near the desired decision boundary.

[Towards Causal VQA
Agarwal et al.]

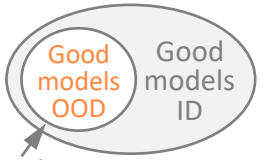
› Training/test sets from different distributions.
Produced by intervening on variable(s) in the data-generating process.

Examples: VQA-CP (intervention on question type & answer), GQA-OOD.



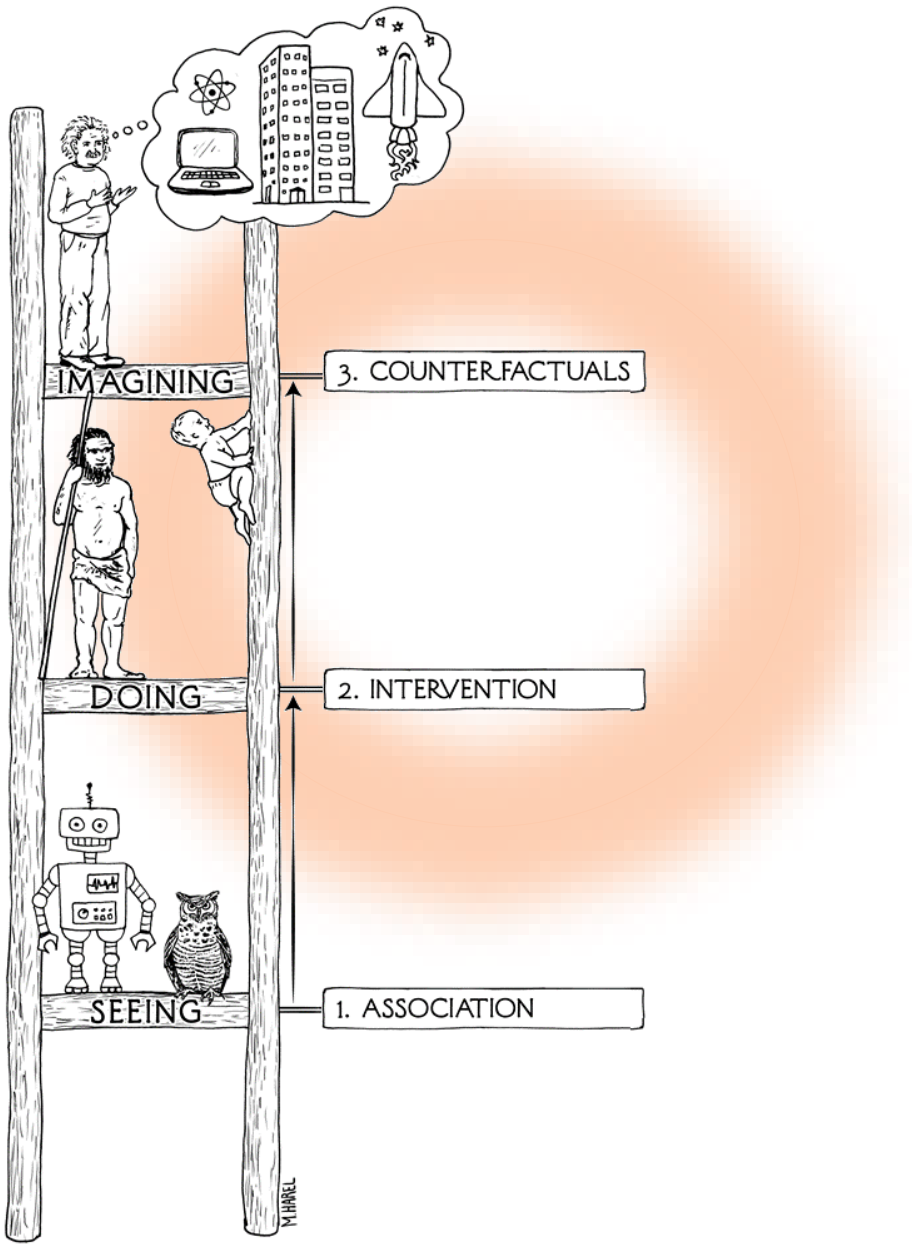
ImageNet-9 backgrounds challenge.
Madry et al.

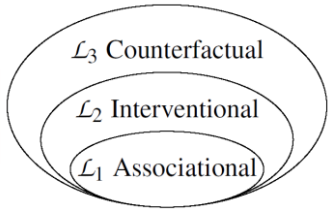
› Classical test set, same distribution as training data.
Measures in-domain (ID) generalization: necessary, not sufficient !
Says nothing about generalization to unusual scenes.



Only a small subset !

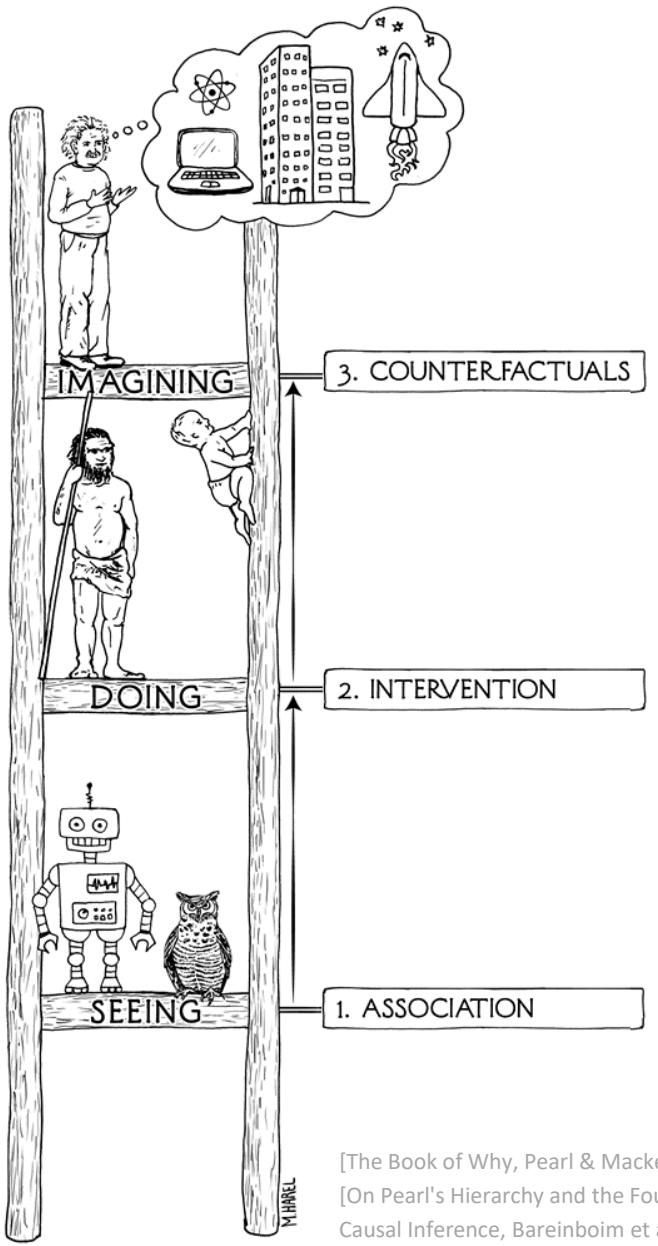
Increasing difficulty →





Each level requires strictly more causal information.
We should design evaluations matching levels 2/3.

Causal hierarchy: 3 types of queries to a model. ⇔ Evaluation settings used in machine learning.



Increasing difficulty →

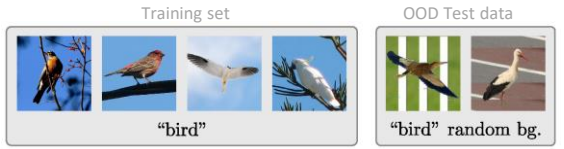
[The Book of Why, Pearl & Mackenzie 2019]
[On Pearl's Hierarchy and the Foundations of Causal Inference, Bareinboim et al. 2021]

› Given a question/image/correct answer, ask the model to generate plausible images supporting alternative answers. Need to understand which visual clues matter.

› Pairs of counterfactual test examples. (a.k.a. contrast sets)
Interventions at instance level. Probe models near the desired decision boundary.

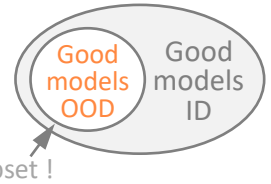
Towards Causal VQA.
Agarwal et al., CVPR 2020.

› Training/test sets from different distributions.
Produced by intervening on variable(s) in the data-generating process.
Examples: VQA-CP (intervention on question type & answer), GQA-OOD.

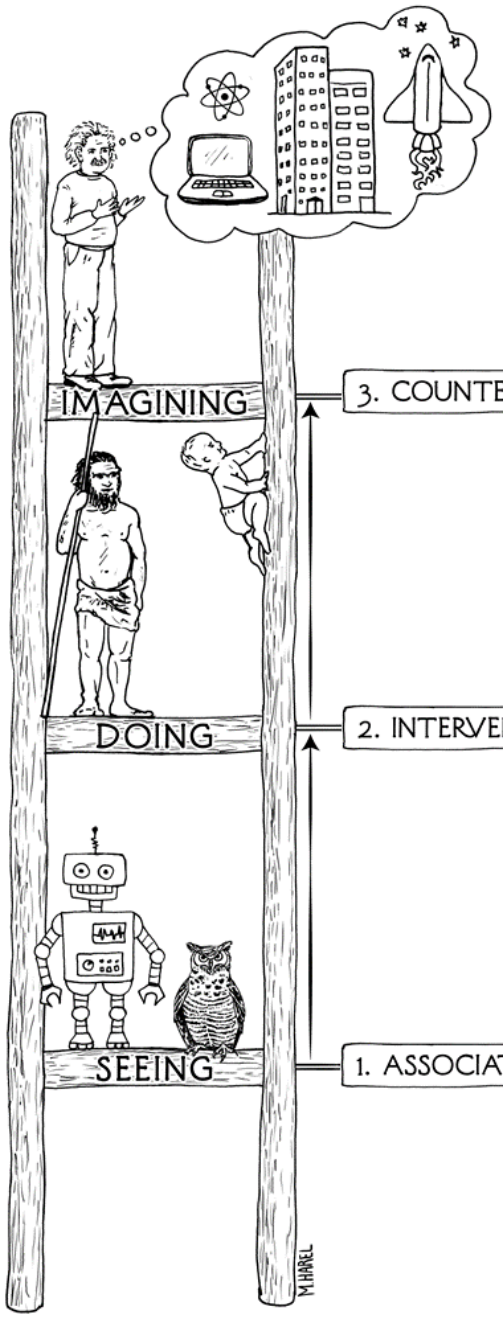


ImageNet-9 backgrounds challenge.
Madry et al.

› Classical test set, same distribution as training data.
Measures in-domain (ID) generalization: necessary, not sufficient !
Says nothing about generalization to unusual scenes.



How to learn *causally-correct* predictive models ?



A model capable of level i requires assumptions/knowledge/data relevant to level $j \geq i$.

⇒ Levels strictly increase in difficulty.

⇒ We cannot learn to reason about interventions (What we really care about.) from observational data (Typical dataset of i.i.d. examples.) alone.

$P(Y_x|x', y')$ $\mathcal{L}_3$ — Counterfactual data
E.g. pairs of counter factual examples.

$P(Y|do(X))$ $\mathcal{L}_2$ — Interventional data
E.g. multiple training distributions.

$P(X, Y)$ $\mathcal{L}_1$ — Observational data
E.g. standard dataset of i.i.d. samples from the joint distribution.

[1] On Pearl's hierarchy and the foundations of causal inference, Bareinboim et al., 2020.

Learning from multiple environments

› One environment = one intervention.

E.g. fixing a selection bias:
pictures from one geographical location, text from one period of time, dialogue from a subgroup of people, sentiment analysis from different domains, ...

Learning from multiple environments

- We want a mapping from e.g. pixels to a representation encoding high-level concepts that causally affect the target.

  Invariant risk minimization (IRM): "To learn invariances across environments, find a data representation such that the optimal classifier on top of that representation matches for all environments."

- The resulting model will make predictions based on features that were invariant across training environments, therefore (under certain conditions) also in new test environments.

- **Some remaining challenges**: optimization difficult, getting data from many diverse environments, trade-off between generalization and lower in-domain performance.

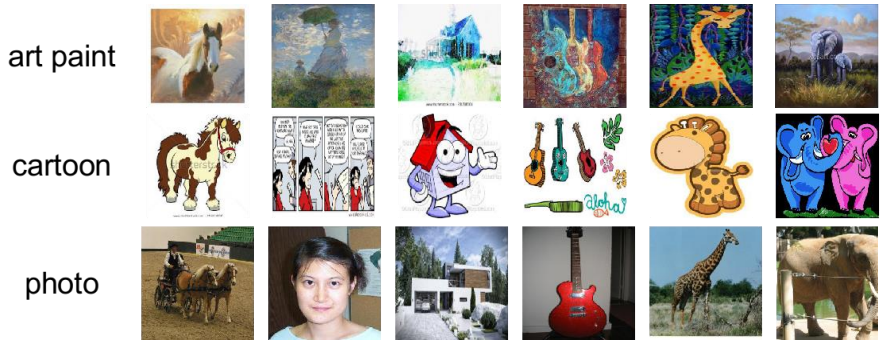
- There are applications of IRM in NLP, e.g. to sentiment classification: OOD generalization is improved by removing spurious reliance on single words that typically correlate highly with the target in the training data.

[1] An Empirical Study of Invariant Risk Minimization, Choe et al. 2020.
[2] Invariant Rationalization, Chang et al., JMLR 2020.

Unshuffling data to improve generalization in VQA

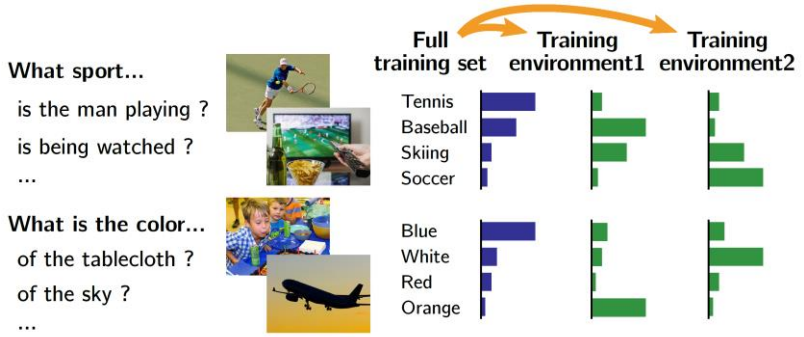
- Learning from multiple training sets / training environments.

Prior work: domain generalization across visual styles [1]



For VQA, we create environments by clustering the training data. [3]

Our work: generalization in VQA



Clustering by question type/answer.

- Intuition: spurious correlations vary across environments, while causal mechanisms remain constant.

Data from environment 1: $(x,y) \sim P_1(X,Y) = P(Y|X)\ \mathbf{P(X|do(Z = z_1))}$

from environment 2: $(x,y) \sim P_2(X,Y) = P(Y|X)\ \mathbf{P(X|do(Z = z_2))}$

Each environment shows an intervention on a variable Z spuriously correlated with labels Y.

- With a well-chosen clustering and modified objective [2] the model is less reliant on answer prior & generalizes better.

Task-specific, human-provided. No free lunch !

[1] Domain generalization – PACS Benchmark, https://domaingeneralization.github.io.

[2] Invariant risk minimization, Arjovsky et al., 2019.

[3] Unshuffling data for improved generalization in visual question answering, Teney et al., CVPR 2021.

Learning from pairs of counterfactual training examples.

Pairs of similar examples with a different label.

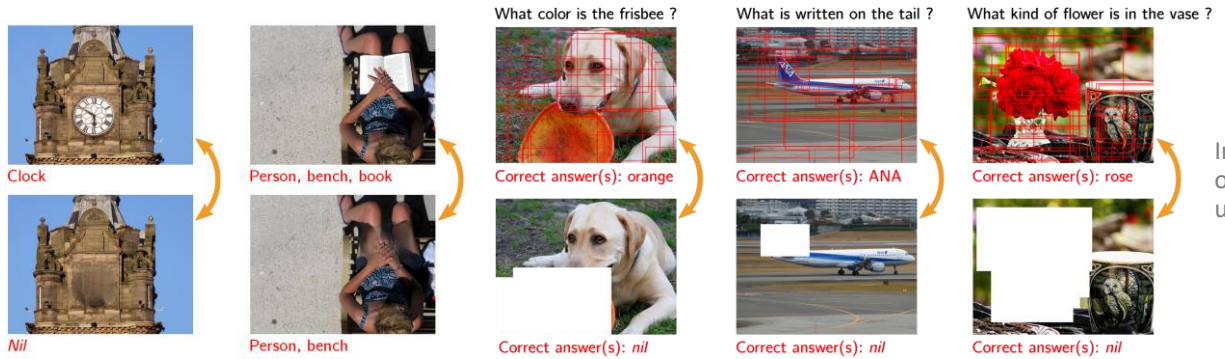
Produced by annotators

Great movie, I loved it. Great editing and use of the soundtrack. **Positive**

Bad movie, I hated it. Weak editing and use of the soundtrack. **Negative**

Minimal edits of movie reviews. [1]

…or generated using heuristics and side annotations.



Inpainting/masking of relevant regions using human attention. [2]

Each pair shows which features (= causal parents) are relevant for flipping the label.
Used as data augmentation: they improve generalization more than the same amount of standard (i.i.d.) data.

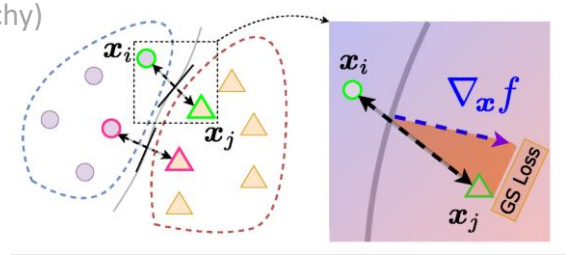
One step further: the causal information is in the relation across each pair. [3] (Level 3 in the causal hierarchy)

We can do better than treating them as individual examples !

New auxiliary loss to exploit these relations.

① Compute vector differences (in feature space) across a pair,

② Align the classifier's gradient (and decision boundary) with it.



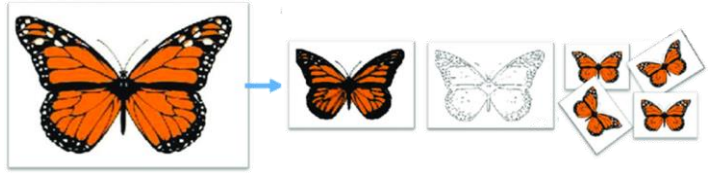
This gives additional improvements in generalization across datasets in VQA, image tagging, textual entailment, sentiment analysis.

[1] Learning the Difference that Makes a Difference with Counterfactually-Augmented Data, Kaushik et al., ICLR 2019.
[2] Towards Causal VQA: Revealing and Reducing Spurious Correlations by Invariant and Covariant Semantic Editing, Agarwal et al., CVPR 2020.
[3] Learning What Makes a Difference from Counterfactual Examples and Gradient Supervision, Teney et al., ECCV 2020.

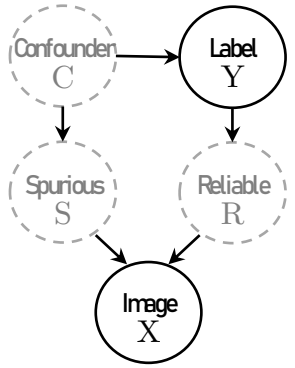
Causal principles help explain the effectiveness of data augmentation and contrastive learning.

Data augmentation simulates interventions.

› **Hard-coded transformations** $(x, y) \rightarrow (x', y)$ produce additional training examples.



› **Causal view**: images contain spurious and reliable features,
both correlated with labels Y because of selection biases (confounder C).
We want a model robust OOD = robust against changes in $P(C)$.



› **This cannot be learned** from samples from the joint $\mathrm{P}(\mathrm{X},\mathrm{Y})$, but it could be learned by observing interventions.
Data augmentation simulates interventions on factors of variation encoded in S. For example:
Augmenting images with geometric transformations = intervening on camera extrinsic parameters.
Augmenting VQA questions with rephrasings = intervening on annotators' writing style.
} *Samples from $P(X, Y|do(S))$, which carry info about causal structure.*

› Root source of improvements = specification of invariances over (X,Y) valid for the (task-specific) data-generating process.
The causal explanation can help select effective augmentations [1].
↘ No universal augmentation !
Similar story in self-supervised/contrastive learning [2]: augmentations = counterfactuals (intervention on style leaving content unchanged).

[1] Selecting Data Augmentation for Simulating Interventions, Ilse et al. ICML 2021.
[2] Self-supervised learning with data augmentations provably isolates content from style, von Kugelgen et al., NeurIPS 2021.

Take-aways

Causality is useful for ML.

- The **capabilities & evaluation of ML models** have outgrown the framework of statistical learning (and its i.i.d. assumption).
- **Causal language helps formalizing existing concepts.**

  E.g. distribution shifts, OOD testing, challenge sets, data augmentation, disentanglement, adversarial examples, ...

  Conditional probabilities cannot describe interventions, causal relationships, or invariances.
- **Causal principles indicate hard limits** on what can be learned from a given type of data and assumptions.

  Comparable to information theory for designing communication systems: not indispensable but darn useful !

When reading papers claiming improved OOD generalization, remember the only 2 possible explanations:

- **Better inductive biases** that make the model closer to the true causal structure of the task.

  Architecture, augmentations, losses, optimizer, ...

  Good accidentally ? Opportunity to discover useful properties of real-world data.

  Good by design ? What are the limits of applicability of the domain knowledge/heuristics used ?
- **Additional training signals** that reveal some of the causal structure of the task.

  I.i.d. samples are not sufficient.

  Alternatives: assumptions (= partial knowledge) about the data-generating process, interventional data (not only w/ RL), multiple training environments, pairs of counterfactual examples, time series, meta data about data collection, …

Open question: how weak (universal) can the assumptions be for causal learning/OOD generalization ?

- Assumptions/heuristics in existing methods are often hidden, and almost surely task/dataset-specific.

  Example: line of works claiming debiasing by removing easy-to-learn features (~ half a dozen paper in the past year).

  The hidden assumption: the second-easiest features are the good ones. Not true in general !

  Still, maybe a useful heuristic in NLP: simple to learn = always spurious ??

  [1] A Too-Good-to-be-True Prior to Reduce Shortcut Reliance, Dagaev et al. 2021.

  [2] Rich Feature Construction for the Optimization-Generalization Dilemma (discussion in Section 5.2), Zhang et al. 2022.

- We cannot do model selection with in-domain validation data (without further assumptions).

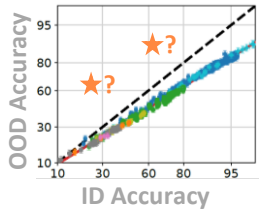
  [3] Evading the simplicity bias: Training a diverse set of models discovers solutions with superior OOD generalization, Teney et al. CVPR 2022.

- OOD Benchmarks can be misused. Example: VQA-CP, many useless papers (still being) published that overfit the OOD test set.

  [4] On the Value of OOD Testing: An Example of Goodhart's Law, Teney et al., NeurIPS 2020.

Open question: how to explain OOD capabilities of large V&L models ?

- Some OOD improvements naturally follow from ID improvements. Effective robustness is rare.



  [5] Accuracy on the Line: On the Strong Correlation Between OOD and ID Generalization, Miller et al. ICML 2021.

  [6] Why do classifier accuracies show linear trends under distribution shift, Mania et al. 2020.

- Effective robustness is lost during fine-tuning.

  [7] The Evolution of OOD Robustness Throughout Fine-Tuning, Andreassen et al., 2021.

- CLIP is exceptionally robust OOD: its data is large and diverse, but also filtered/selected/weighted.

  [8] Data Determines Distributional Robustness in CLIP, Fang et al., 2022.

Further reading

Improving ML with causality (introductions and reviews):

- *Causality for Machine Learning*, Cloudera report, 2020.
- *Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond*, Feder at al., 2021.
- *Towards Causal Representation Learning*, Scholkopf et al. 2021.
- *From Statistical to Causal Learning*, Scholkopf and Kuegelgen, 2022.
- *Causality matters in medical imaging*, Castro et al., Nature Communications, 2020.

Not covered here: relevance of causality to fairness:

- *Causal Reasoning for Algorithmic Fairness*, Loftus et al., 2018.
- *Avoiding Discrimination through Causal Reasoning*, Kilbertus et al., NeurIPS 2017.
- *On the Fairness of Disentangled Representations*, Locatello et al., NeurIPS 2019.

Also not covered here: improving causal inference (answering causal questions) with ML:

- *Causal Effects of Linguistic Properties*, Pryzant et al., 2021.

Extra slides

Advantages of a causal model

- **The causal structure of the data-generating process is more informative than statistical information.**

  P(B|A) *Conditional distribution* = *Filtering an observed distribution.*

  ≠

  P(B|do(A)) *Interventional distribution* = *Forcing a variable to a specific value.*

- **Learning causal structure = learning invariants.**

  Causal relationships hold true across environments, by definition.

  They allow predicting the effect of interventions in conditions not seen during training.

- **Always preferable to a statistical model ? No.**

  If training/test distributions are similar, predictions can be better/easier using all correlations (incl. spurious ones).

  Because causal correlations being more noisy/difficult to learn (than spurious ones).

  E.g. if red cars are always fast and cows are always on green grass (and vice versa!), predictions are easy using just color!

Do we always want to rely only on causal features ?

- No, we can safely use contextual cues when training/test distributions are guaranteed to stay similar.

  E.g. use background for recognition.

- Trade-off: greater predictive accuracy vs. better generalization to distribution shifts.
- It's important to make these assumptions & choices explicit.

  E.g. for medical imaging , in high-stakes ML, for understanding/eliminating unfair biases.