Visual question answering and reasoning over vision and language.

Beyond the limits of statistical learning ?

Damien Teney

Idiap Research Institute, Martigny, Switzerland.

Australian Institute for Machine Learning, Adelaide, Australia.

MSR Seminar on Vision and Language

August 2021


Martigny, Switzerland
https://idiap.ch/jobs
idiap
RESEARCH INSTITUTE

Visual question answering (VQA) is exciting because it's a general, complex task.

What is the mustache made of ?

....especially in view of the relevant (?) training examples. (banana- and moustache-related samples from VQA v2)

What is presented to the winner ?
Ground truth answer(s): bananas.

What is this person listening to ?
GT: banana.

What color is the gentleman's mustache ?
GT: gray, silver.

What is his mustache ?
GT: hair, fake, don't know, handlebar.

VQA requires out-of-distribution (OOD) generalization.

= Applying learned concepts & reasoning mechanisms beyond the training distribution.

Unsolved problem, even on toy data. 

Empirical risk minimization (ERM) = learning by association, of any correlation between inputs/labels.

Training data

~~"bird"~~

"Sunny day and tree branches"

→ OOD Generalization is underspecified by this data !

OOD Test data

"bird" random bg.

ImageNet-9 backgrounds challenge, Madry lab. MIT.

Classical in-domain generalization

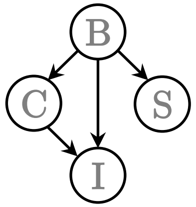
- Means "filling the gaps" between training examples.
- Generally useful inductive biases (smoothness, Occam's razor).
- More data helps (solved with infinite data).

Strong out-of-domain generalization

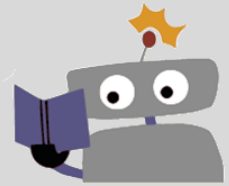
- Means distinguishing "robust" vs "spurious" features.
- Requires additional (task-specific) knowledge.
- More (of the same, biased) data **does NOT help** !

Complex tasks like VQA require more than classical statistical learning & learning by association.

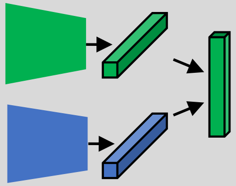
Some causal principles



Implications for evaluation



Implications for learning



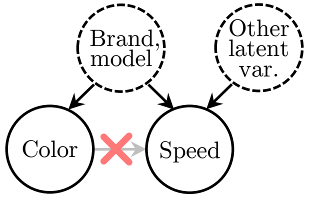
Why does causality matter ?

X causes Y $X \rightarrow Y$ $\Leftrightarrow$ Intervening on X affects Y.

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

Example task: predicting the top speed of a car from an image.

Training images annotated with speed



Probability that a car of a certain color can go fast.

A statistical model learns correlations. "red = fast"

Reliable only if training/test data are from the same distribution.

$$\mathrm{P}(\mathit{Speed} \mid \mathit{Color}) \neq \mathrm{P}(\mathit{Speed} \mid \mathrm{do}(\mathit{Color}))$$

A causal model encodes the effects of interventions.

Enables predictions in conditions unobserved during training (i.e. OOD).

Probability that a car can go fast after being re-painted in a certain color.

What would happen to a re-painted car ?

→

Faster ? No !

More (of the same) **data does not improve the statistical model OOD !**

More red Ferraris don't help distinguishing spurious correlations from causal mechanisms.

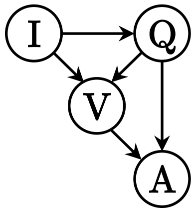
Back to VQA...

Real world | Learned model

Real world

- A set of mechanisms produce the observed (training) data.
- Its causal structure defines which variables/features are correlated/robust/spurious.

The data-generating process in VQA is **a human annotator**
who takes an Image and Question as input,
finds relevant Visual information,
and produces an Answer.



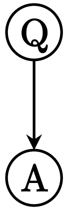
$$P(I,Q,V,A) = \mathbf{P(A|V,Q)}\ \mathbf{P(V|I,Q)}\ P(Q|I)\ P(I)$$

Mechanisms guaranteed to transfer out-of-distribution.

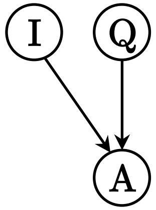
Learned models

- Trained to mimic the real world.
- The inference also has a causal structure: inputs → … → predictions.

A bad VQA model that guesses answers without looking at the image.



A VQA model with no attention mechanism.



We want to mirror some of the causal structure & mechanisms.

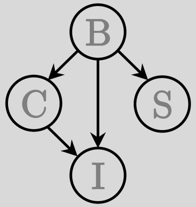
Why is it hard ? Because this information is absent from typical datasets !

I.i.d. training samples (observational data) are generally insufficient to recover the causal structure.

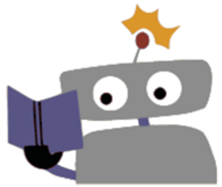
We need additional assumptions, or task-specific knowledge about the causal structure, or other types of data.

(e.g. as inductive biases like attention architectures)

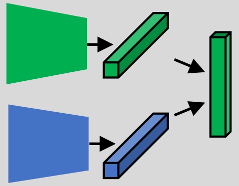
Causal principles



Implications for evaluation

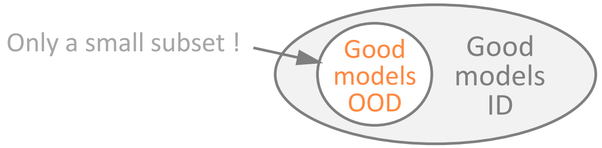


Implications for learning

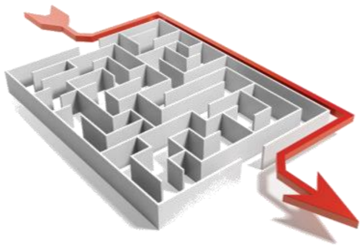


Can the model answer arbitrary questions about novel unusual scenes ?

- A classical test set measures in-domain (ID) generalization (same distribution as the training set).
- ID performance is necessary but not sufficient !

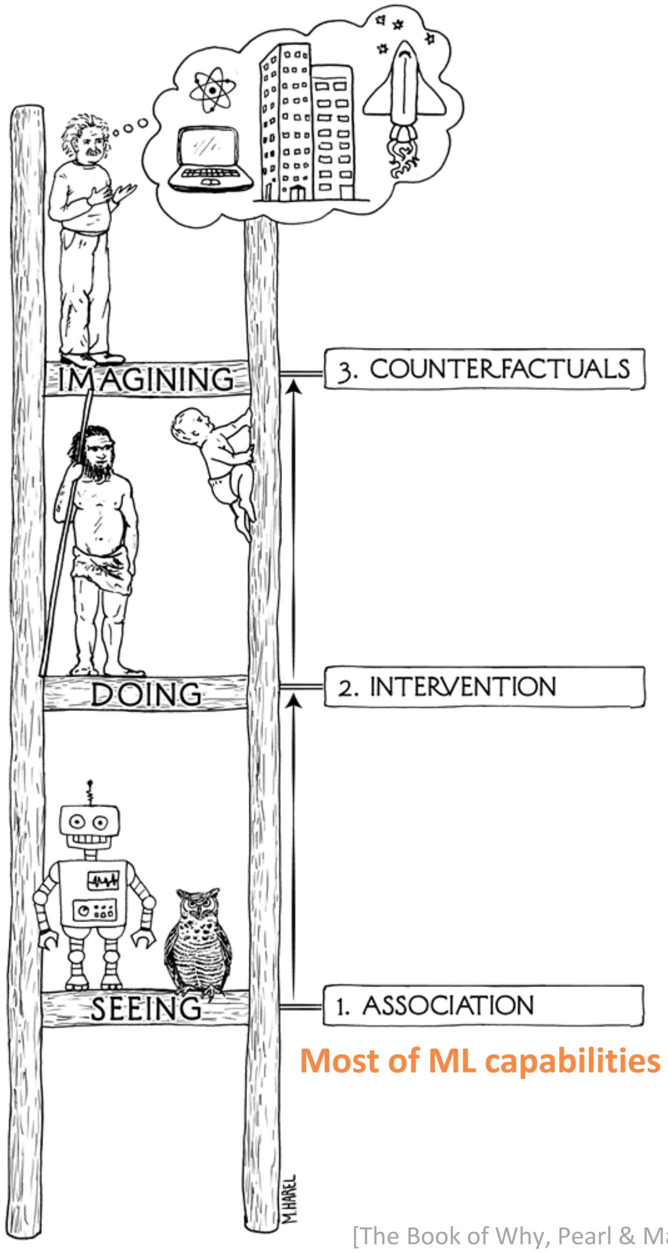


- ID performance says nothing about OOD generalization and reliance on shortcut learning. Sky=bird, red=fast, question biases…



- We've already made a lot of progress: VQA v2, VQA-CP, GQA-OOD, counterfactual examples, etc. These can be formalized with causal principles.

Judea Pearl's causal hierarchy defines three types of queries of increasing difficulty we can make to a model.



Most of ML capabilities today.

[The Book of Why, Pearl & Mackenzie 2019]

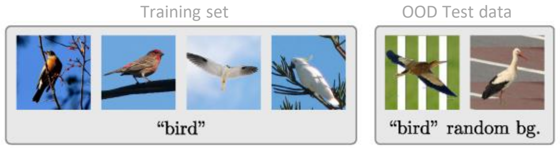
Pairs of counterfactual examples ≈ intervention at instance level.
Intuitively, we probe the model close to the desired decision boundary.

Examples: VQA v2 (balanced pairs), [Towards Causal VQA], [Evaluating NLP Models via Contrast Sets], [Automatic Generation of Contrast Sets from Scene Graphs].

An even harder idea: require the inverted model to generate plausible images for alternative answers.

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

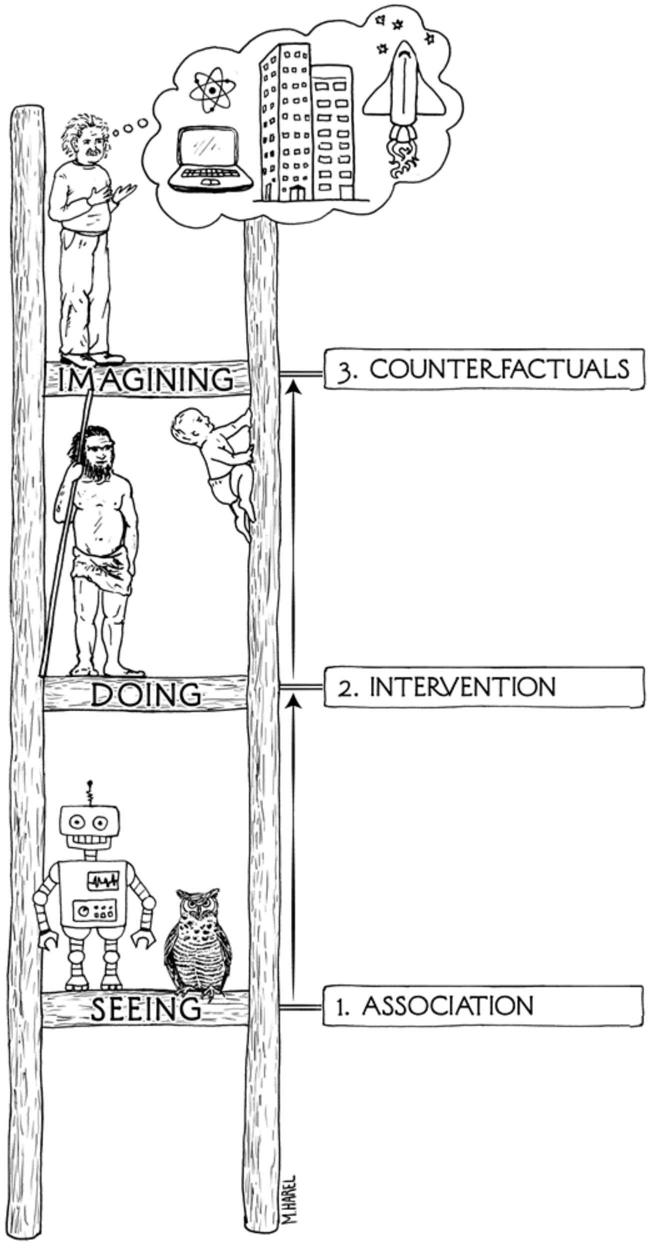
Examples: VQA-CP (intervention on question type & answer), GQA-OOD.

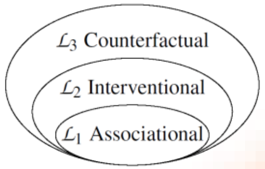


ImageNet-9 backgrounds challenge, Madry lab, MIT.

Classical test set from the same distribution as the training data.
Cannot measure OOD generalization.

Examples: VQA v1, GQA.





Each level requires strictly more causal information.

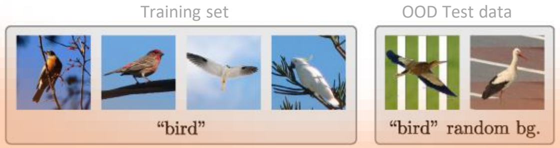
Pairs of counterfactual examples ≈ intervention at instance level.
Intuitively, we probe the model close to the desired decision boundary.

Examples: VQA v2 (balanced pairs), [Towards Causal VQA], [Evaluating NLP Models via Contrast Sets], [Automatic Generation of Contrast Sets from Scene Graphs].

An even harder idea: require the inverted model to generate plausible images for alternative answers.

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

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

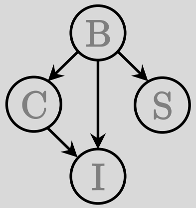


ImageNet-9 backgrounds challenge. (Madry lab, MIT)

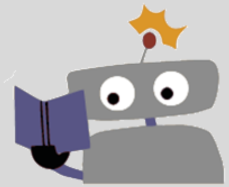
Classical test set from the same distribution as the training data.
Cannot measure OOD generalization.

Examples: VQA v1, GQA.

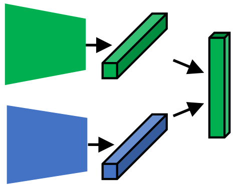
Causal principles



Implications for evaluation



Implications for learning



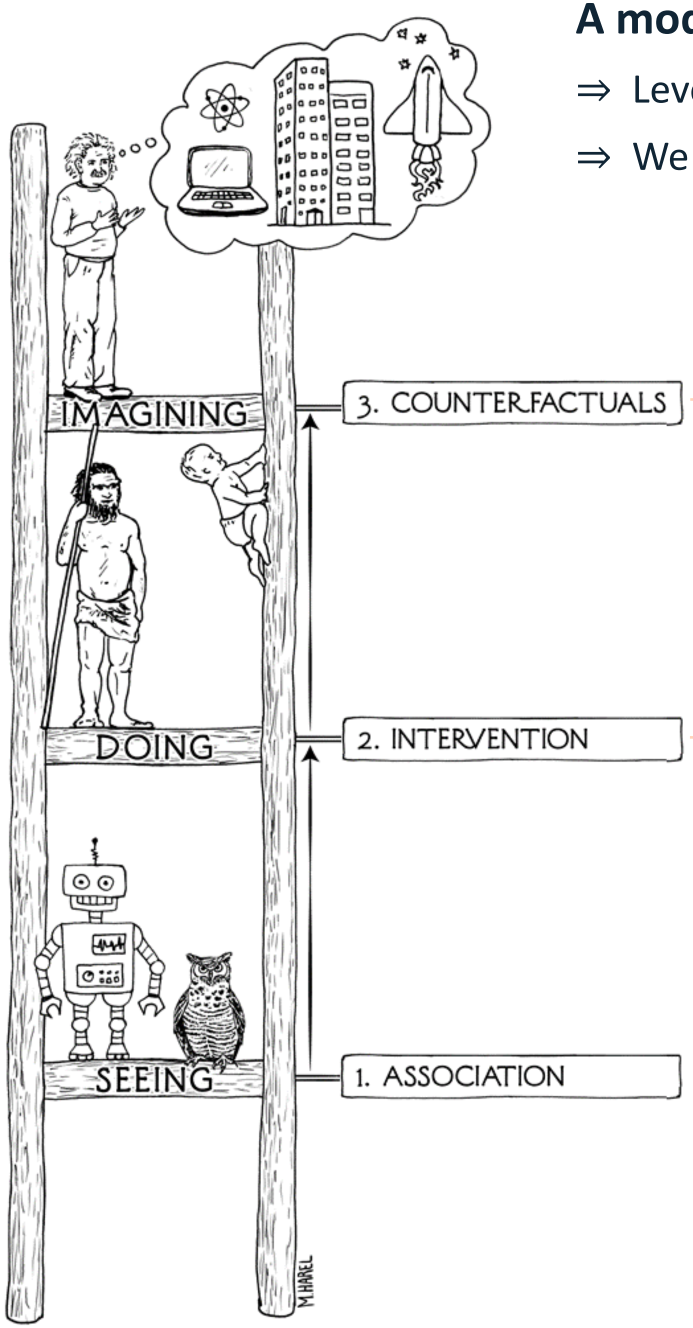
A model capable of level i requires assumptions/knowledge/data relevant to level $j \geq i$.

⇒ Levels strictly increase in difficulty.

What we really care about.

Typical dataset of i.i.d. examples.

⇒ We cannot learn to reason about interventions from observational data 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.

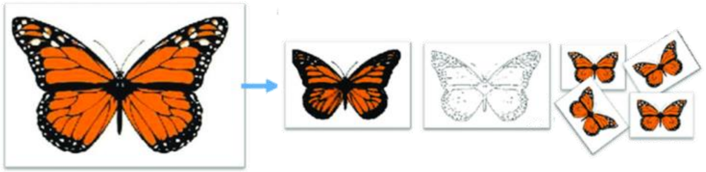
$P(X, Y)$
$\mathcal{L}_1$

Observational data
E.g. standard dataset of i.i.d. samples from the joint distribution.

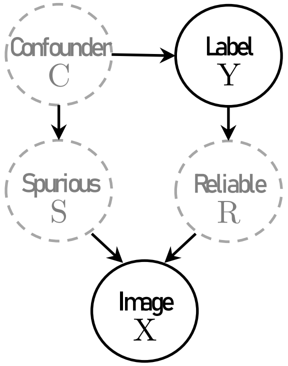
[On Pearl's hierarchy and the foundations of causal inference, Bareinboim et al. 2020]

We can explain successful techniques from a causal perspective.

Data augmentation simulates interventions.



- Hard-coded transformations $(x, y) \to (x', y')$ into additional training examples.



- Images contain spurious and reliable features, both correlated with labels Y because of hidden confounders C.
  We want a model robust OOD i.e. robust against changes in $\mathrm{P(C)}$.

- This cannot be learned from samples from the joint $\mathrm{P(X,Y)}$ but it could be learned by observing interventions (level-2 information).
  Data augmentation simulates interventions by editing (spurious) factors in variation encoded in S.
  - Augmenting images with geometric transformations = intervening on camera extrinsic parameters.
  - Augmenting VQA questions with rephrasings = intervening on annotators' writing style.

  } Samples from $\mathrm{P(X,Y \mid do(S))}$. Carry info about causal mechanisms.

- **What did we learn ?**
  The root source of improvement = specification of invariances over $\mathrm{(X,Y)}$ that are valid for the (task-specific) data-generating process.
  No universal augmentation !
  Can also help select effective augmentations: Selecting Data Augmentation for Simulating Interventions, Ilse et al. ICML 2021.

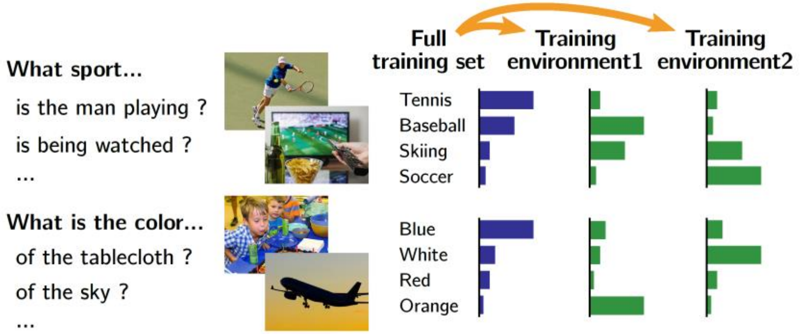
Unshuffling data recovers non-i.i.d. subsets of training data

[Unshuffling Data for Improved Generalization in VQA, Teney et al. ICCV 2021]

> **Existing work.** Domain generalization, data collected in multiple conditions.
Various methods can learn a predictor robust across environments. [ICP, IRM, ...]

PACS Dataset.

Our method. Cluster non-i.i.d. subsets, using task knowledge & side annotations.



Clustering VQA-CP by question type/answer.

> **The causal perspective:** we have more information than the aggregated (i.i.d.) data.

With each cluster, we observe an intervention on a variable $Z \not\perp\!\!\!\perp Y$, spuriously correlated with labels Y but not a direct causal parent. $Z \rightarrow X \rightarrow Y$

We make the spurious correlations vary across clusters. We know that the causal mechanisms (to learn) stay invariant. [Principle of independent mechanisms, Peters 2017]

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

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

> Root source of improvement = the well-chosen (task-specific) clustering condition.

No free lunch !

Counterfactual training examples provide level-3 information.

[Learning What Makes a Difference from Counterfactual Examples, Teney et al. ECCV 2020]

Pairs of similar examples with a different label.

Produced by annotators

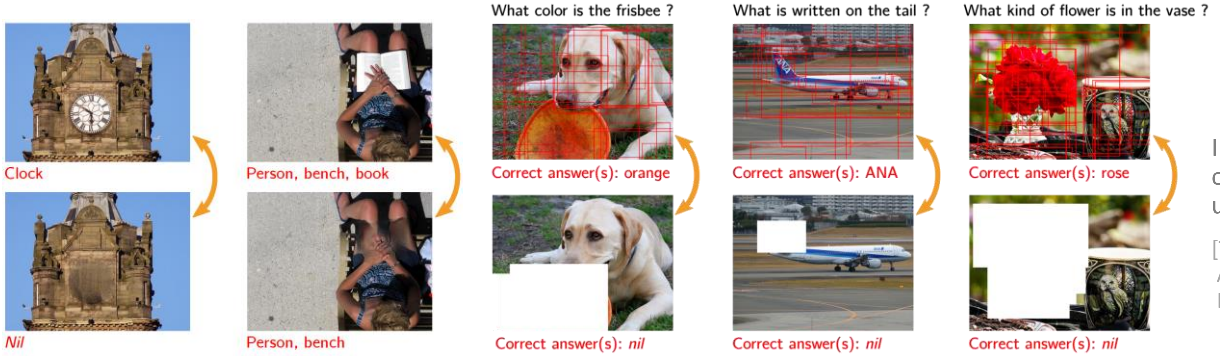
Great movie, I loved it. Great editing and use of the soundtrack. **Positive**

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

Minimal edits of movie reviews.

[Learning the diff, Kaushik et al. 2019]

...or generated using (task-specific !) heuristics and side annotations.



Inpainting/masking of relevant regions using human attention.

[Towards Causal VQA, Agarwal et al. 2020; MUTANT, Gokhale et al. 2020]

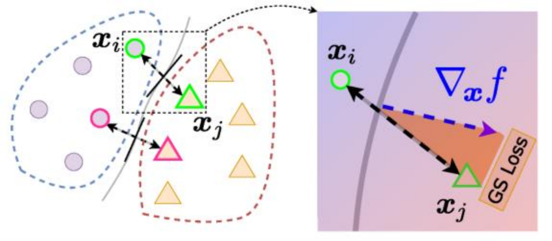
Each pair shows which features are relevant to flip the label (= causal parents) ⇒ They improve generalization more than the same amount of standard i.i.d. data.

The causal perspective: the level-3 causal information is in the relation across each pair.

We can do better than treating them as individual examples !

We designed a loss to exploit the relation. 1 Compute vector differences (in feature space) across a pair.

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



We get additional improvements in generalization across datasets in VQA, image tagging, textual entailment, sentiment analysis.

Takeaways

There are fundamental limits to what can be learned from i.i.d. training examples, no matter how many.

- Causal principles set hard limits on which properties of the world can be learned in given training conditions.

  [On Pearl's Hierarchy and the Foundations of Causal Inference, Bareinboim et al. 2021]

You would like to learn a causally-accurate model, even if you don't know it.

- It ensures generalization to arbitrary (covariate) distribution shifts.
- Causal principles may not directly inform the design of learning algorithms… but they point at sources for finding the missing information.

~~Fixed training sets~~
~~(iid samples from a joint dist.)~~

→

Training environments

- Data collected in multiple conditions
- Meta annotations e.g. annotator identity
- Non-stationary data
- Interventional data
- Counterfactual examples   Etc.

The road ahead.

[Towards Causal Representation Learning, Scholkopf et al. 2021]
[Inductive Biases for Deep Learning of Higher-Level Cognition, Goyal and Bengio 2021]