Causal Representation Learning

SML Journal Club

Emanuele Marconato

University of Pisa, Department of Computer Science
University of Trento, Department of Information Engineering and Computer Science

About us: SML Journal Club.

The Journal Club structure is still under discussion:

- Open to all interested people;
- Keep blended modality;
- We are planning to have invited speakers;
- Once every two weeks, on Friday afternoon at 16:00 CET.

$\rightarrow$ You are welcome to have a drink with us after the talk session $\leftarrow$

Table of Contents

Towards Causal Representation Learning

Bernhard Schölkopf [†], Francesco Locatello [†], Stefan Bauer [*], Nan Rosemary Ke [*], Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

Table of Contents

Towards Causal Representation Learning

Bernhard Schölkopf [†], Francesco Locatello [†], Stefan Bauer [*], Nan Rosemary Ke [*], Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

Table of Contents

Towards Causal Representation Learning

Bernhard Schölkopf [†], Francesco Locatello [†], Stefan Bauer [*], Nan Rosemary Ke [*], Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

Table of Contents

Towards Causal Representation Learning

Bernhard Schölkopf [†], Francesco Locatello [†], Stefan Bauer [*], Nan Rosemary Ke [*], Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

Table of Contents

Towards Causal Representation Learning

Bernhard Schölkopf [†], Francesco Locatello [†], Stefan Bauer [*], Nan Rosemary Ke [*], Nal Kalchbrenner
Anirudh Goyal, Yoshua Bengio

Motivation

Limitations of Machine Learning:

- It is rooted on the $I.I.D.$ hypothesis and does not work well outside it;
- It is weak against noises and confounders;
- It is not reusable;
- It does not allow any knowledge beyond typically statistical reasoning.

On the other hand, we (humans) acquire knowledge by:

- Understanding the relevant information, even in noisy contexts;
- Being able to generalize outside the distribution;
- We can infer causal, or physical, models out of our observations: learning transferable knowledge to other domains.

We focus on learning useful representations of data.

Motivation

Limitations of Machine Learning:

- It is rooted on the $I.I.D.$ hypothesis and does not work well outside it;
- It is weak against noises and confounders;
- It is not reusable;
- It does not allow any knowledge beyond typically statistical reasoning.

On the other hand, we (humans) acquire knowledge by:

- Understanding the relevant information, even in noisy contexts;
- Being able to generalize outside the distribution;
- We can infer causal, or physical, models out of our observations: learning transferable knowledge to other domains.

We focus on learning useful representations of data.

Motivation

Limitations of Machine Learning:

- It is rooted on the *I.I.D.* hypothesis and does not work well outside it;
- It is weak against noises and confounders;
- It is not reusable;
- It does not allow any knowledge beyond typically statistical reasoning.

On the other hand, we (humans) acquire knowledge by:

- Understanding the relevant information, even in noisy contexts;
- Being able to generalize outside the distribution;
- We can infer causal, or physical, models out of our observations: learning transferable knowledge to other domains.

We focus on learning useful representations of data.

Motivation

Limitations of Machine Learning:

- It is rooted on the $I.I.D.$ hypothesis and does not work well outside it;
- It is weak against noises and confounders;
- It is not reusable;
- It does not allow any knowledge beyond typically statistical reasoning.

On the other hand, we (humans) acquire knowledge by:

- Understanding the relevant information, even in noisy contexts;
- Being able to generalize outside the distribution;
- We can infer causal, or physical, models out of our observations: learning transferable knowledge to other domains.

We focus on learning useful representations of data.

Motivation

Limitations of Machine Learning:

- It is rooted on the $I.I.D.$ hypothesis and does not work well outside it;
- It is weak against noises and confounders;
- It is not reusable;
- It does not allow any knowledge beyond typically statistical reasoning.

On the other hand, we (humans) acquire knowledge by:

- Understanding the relevant information, even in noisy contexts;
- Being able to generalize outside the distribution;
- We can infer causal, or physical, models out of our observations: learning transferable knowledge to other domains.

We focus on learning useful representations of data.

Motivation

Limitations of Machine Learning:

- It is rooted on the $I.I.D.$ hypothesis and does not work well outside it;
- It is weak against noises and confounders;
- It is not reusable;
- It does not allow any knowledge beyond typically statistical reasoning.

On the other hand, we (humans) acquire knowledge by:

- Understanding the relevant information, even in noisy contexts;
- Being able to generalize outside the distribution;
- We can infer causal, or physical, models out of our observations: learning transferable knowledge to other domains.

We focus on learning useful representations of data.

Motivation

Limitations of Machine Learning:

- It is rooted on the $I.I.D.$ hypothesis and does not work well outside it;
- It is weak against noises and confounders;
- It is not reusable;
- It does not allow any knowledge beyond typically statistical reasoning.

On the other hand, we (humans) acquire knowledge by:

- Understanding the relevant information, even in noisy contexts;
- Being able to generalize outside the distribution;
- We can infer causal, or physical, models out of our observations: learning transferable knowledge to other domains.

We focus on learning useful representations of data.

Motivation

Limitations of Machine Learning:

- It is rooted on the $I.I.D.$ hypothesis and does not work well outside it;
- It is weak against noises and confounders;
- It is not reusable;
- It does not allow any knowledge beyond typically statistical reasoning.

On the other hand, we (humans) acquire knowledge by:

- □ Understanding the relevant information, even in noisy contexts;
- ◎ Being able to generalize outside the distribution;
- △ We can infer causal, or physical, models out of our observations: learning transferable knowledge to other domains.

We focus on learning useful representations of data.

Motivation

Limitations of Machine Learning:

- It is rooted on the $I.I.D.$ hypothesis and does not work well outside it;
- It is weak against noises and confounders;
- It is not reusable;
- It does not allow any knowledge beyond typically statistical reasoning.

On the other hand, we (humans) acquire knowledge by:

- □ Understanding the relevant information, even in noisy contexts;
- ◎ Being able to generalize outside the distribution;
- △ We can infer causal, or physical, models out of our observations: learning transferable knowledge to other domains.

We focus on learning useful representations of data.

Representation Learning (2013)

Addressing the learning of **factors of variation** by extracting a useful representation of input data:

$$\mathbf{x} \rightarrow r(\mathbf{x}) \quad r : \mathbb{R}^d \rightarrow \mathbb{R}^n.$$

Hypothesis: the intractable input distribution $p(\mathbf{x})$ is originated from a simpler *latent* distribution $p(\mathbf{z})$, such that:

$$p(\mathbf{x}) = \int d\mathbf{z}\, p(\mathbf{z}) p(\mathbf{x}|\mathbf{z})$$

where in Representation Learning $p(\mathbf{x}|\mathbf{z}) := p_\theta(\mathbf{x}|\mathbf{z})$ is a conditional distribution "decoding" the latent factors to the sensorial inputs.

Representation Learning (2013)

Addressing the learning of **factors of variation** by extracting a useful representation of input data:

$$\mathbf{x} \rightarrow r(\mathbf{x}) \quad r : \mathbb{R}^d \rightarrow \mathbb{R}^n.$$

Hypothesis: the intractable input distribution $p(\mathbf{x})$ is originated from a simpler *latent* distribution $p(\mathbf{z})$, such that:

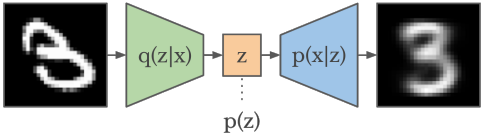
$$p(\mathbf{x}) = \int d\mathbf{z}\, p(\mathbf{z}) p(\mathbf{x}|\mathbf{z})$$

where in Representation Learning $p(\mathbf{x}|\mathbf{z}) := p_\theta(\mathbf{x}|\mathbf{z})$ is a conditional distribution "decoding" the latent factors to the sensorial inputs.

VAE and Disentangled Factors of Variations

It is generically difficult to understand useful latent representations of data. We divide the the problem in two pieces:

- **Encoding** $q_\phi(\mathbf{z}|\mathbf{x})$, e.g. Convolutional NN.
- **Decoding** $p_\theta(\mathbf{x}|\mathbf{z})$.



But when factors of variation $\mathbf{z}$ are useful? They must be independent and represent a dimension over one change occurs: in that case they are **disentangled**.

A Variational Auto-Encoder (VAE) forces the learning of useful representation by pushing the encoder distribution to the chosen prior $p(\mathbf{z})$, e.g.:

$$p(\mathbf{z}) = \prod_i p(\mathbf{z}_i)$$

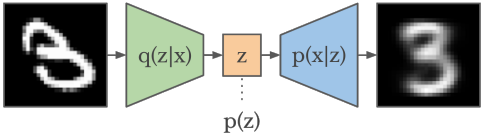
promotes the learning of independent latent dimensions.

However this works only in few idealized cases, and inferring a suitable prior for $\mathbf{z}$ is referred as the **prior hole** problem!

VAE and Disentangled Factors of Variations

It is generically difficult to understand useful latent representations of data. We divide the the problem in two pieces:

- **Encoding** $q_\phi(\mathbf{z}|\mathbf{x})$, e.g. Convolutional NN.
- **Decoding** $p_\theta(\mathbf{x}|\mathbf{z})$.



But when factors of variation $\mathbf{z}$ are useful? They must be independent and represent a dimension over one change occurs: in that case they are **disentangled**.

A Variational Auto-Encoder (VAE) forces the learning of useful representation by pushing the encoder distribution to the chosen prior $p(\mathbf{z})$, e.g.:

$$p(\mathbf{z}) = \prod_i p(\mathbf{z}_i)$$

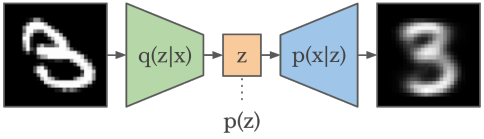
promotes the learning of independent latent dimensions.

However this works only in few idealized cases, and inferring a suitable prior for $\mathbf{z}$ is referred as the **prior hole** problem!

VAE and Disentangled Factors of Variations

It is generically difficult to understand useful latent representations of data. We divide the the problem in two pieces:

- **Encoding** $q_\phi(\mathbf{z}|\mathbf{x})$, e.g. Convolutional NN.
- **Decoding** $p_\theta(\mathbf{x}|\mathbf{z})$.



But when factors of variation $\mathbf{z}$ are useful? They must be independent and represent a dimension over one change occurs: in that case they are **disentangled**.

A Variational Auto-Encoder (VAE) forces the learning of useful representation by pushing the encoder distribution to the chosen prior $p(\mathbf{z})$, e.g.:

$$p(\mathbf{z}) = \prod_i p(\mathbf{z}_i)$$

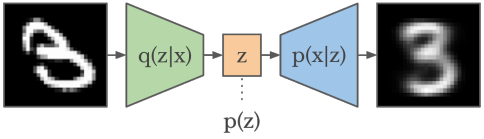
promotes the learning of independent latent dimensions.

However this works only in few idealized cases, and inferring a suitable prior for $\mathbf{z}$ is referred as the **prior hole** problem!

VAE and Disentangled Factors of Variations

It is generically difficult to understand useful latent representations of data. We divide the the problem in two pieces:

- **Encoding** $q_\phi(\mathbf{z}|\mathbf{x})$, e.g. Convolutional NN.
- **Decoding** $p_\theta(\mathbf{x}|\mathbf{z})$.



But when factors of variation $\mathbf{z}$ are useful? They must be independent and represent a dimension over one change occurs: in that case they are **disentangled**.

A Variational Auto-Encoder (VAE) forces the learning of useful representation by pushing the encoder distribution to the chosen prior $p(\mathbf{z})$, e.g.:

$$p(\mathbf{z}) = \prod_i p(\mathbf{z}_i)$$

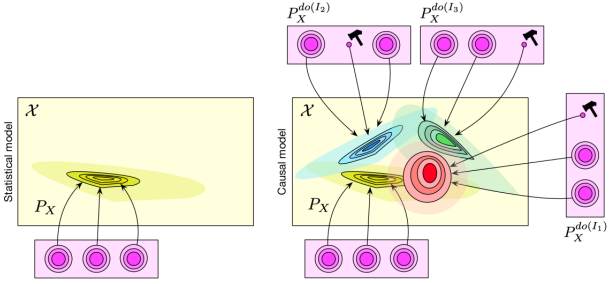
promotes the learning of independent latent dimensions.

However this works only in few idealized cases, and inferring a suitable prior for $\mathbf{z}$ is referred as the **prior hole** problem!

Levels of Causal Modelling

What can we learn out of data?

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes



- Interventions over causal factors lead to a drift in the observed distribution.
- Causal relations contain more information than statistical ones.

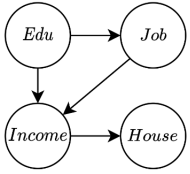
Structural Causal Models

Let $\mathcal{X} = \{X_i\}_{i=1}^n$ a set of observables that form a direct acyclic graph (DAG):

$$X_i = f_i(\mathbf{PA}_i, U_i), \forall i \implies P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|\mathbf{PA}_i)$$

The probabilistic decomposition of $P(X_1, ..., X_n)$, given the DAG, is the causal factorization of the ensemble $\mathcal{X}$. A Structural Causal model, also explains:

- **Interventions**
- **Counterfactuals**



a_1 : get_degree(bachelor)
a_2 : change_job(developer)
a_3 : change_house(buy)

Indipendent Causal Mechanism Principle

The decomposition of the Structural Causal Model implies a structure of statistical independence among variables ($i \neq j$):

$$P(X_i|\mathbf{PA}_i) \perp\!\!\!\perp P(X_j|\mathbf{PA}_j)$$

1. **no influence:** changing one mechanism $P(X_i|\mathbf{PA}_i)$ does not change other mechanisms $P(X_j|\mathbf{PA}_j)$;
2. **no information:** knowing some other mechanisms $P(X_i|\mathbf{PA}_i)$ does not give us information about a mechanism $P(X_j|\mathbf{PA}_j)$.

Indipendent Causal Mechanism Principle

The decomposition of the Structural Causal Model implies a structure of statistical independence among variables ($i \neq j$):

$$P(X_i|\mathbf{PA}_i) \perp\!\!\!\perp P(X_j|\mathbf{PA}_j)$$

1. **no influence:** changing one mechanism $P(X_i|\mathbf{PA}_i)$ does not change other mechanisms $P(X_j|\mathbf{PA}_j)$;
2. **no information:** knowing some other mechanisms $P(X_i|\mathbf{PA}_i)$ does not give us information about a mechanism $P(X_j|\mathbf{PA}_j)$.

Indipendent Causal Mechanism Principle

The decomposition of the Structural Causal Model implies a structure of statistical independence among variables ($i \neq j$):

$$P(X_i|\mathbf{PA}_i) \perp\!\!\!\perp P(X_j|\mathbf{PA}_j)$$

1. **no influence:** changing one mechanism $P(X_i|\mathbf{PA}_i)$ does not change other mechanisms $P(X_j|\mathbf{PA}_j)$;
2. **no information:** knowing some other mechanisms $P(X_i|\mathbf{PA}_i)$ does not give us information about a mechanism $P(X_j|\mathbf{PA}_j)$.

Machine Learning and Causality

Challenges:

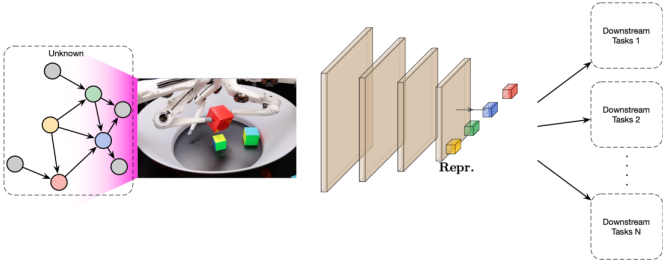
- Infer causal variables from the available low-level input features;
- There is no consensus on which aspects of the data reveal causal relations.

Learning Causal Representations

Learning disentangled representation of causal variables, $\mathbf{e} : \mathbb{R}^d \to \mathbb{R}^n$ with $n \ll d$:

$$z_i = f_i(\mathbf{PA}_i, U_i) \quad (i = 1, ..., n)$$

but $\mathbf{PA}_i = \emptyset, \ \forall i$. In practice a decoder $\mathbf{d} = p \circ f$, learns a hierarchy of disentangled factors [2].
This depends on which *interventions* we observe $\implies$ shift from usual i.i.d datasets [3]!

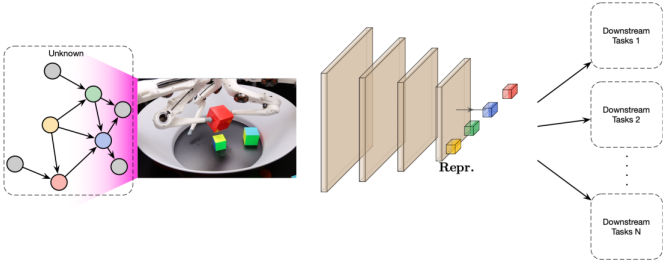


Learning Causal Representations

Learning disentangled representation of causal variables, $\mathbf{e} : \mathbb{R}^d \to \mathbb{R}^n$ with $n \ll d$:

$$z_i = f_i(\mathbf{PA}_i, U_i) \quad (i = 1, ..., n)$$

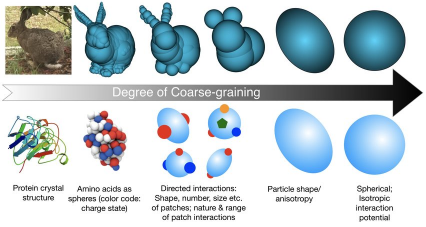
but $\mathbf{PA}_i = \emptyset, \ \forall i$. In practice a decoder $\mathbf{d} = p \circ f$, learns a hierarchy of disentangled factors [2].
This depends on which *interventions* we observe $\implies$ shift from usual i.i.d datasets [3]!



Learning Causal Representations

A priori, we must face the problem of what **representation** preserves the causal structure hidden on observations $\mathbf{X}$.

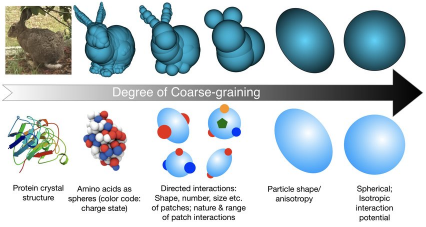
- Understand those *coarse-graining* maps preserving important relations, [4].
- Discover the conditional dependence among latents factors z_i, [5].



Learning Causal Representations

A priori, we must face the problem of what **representation** preserves the causal structure hidden on observations **X**.

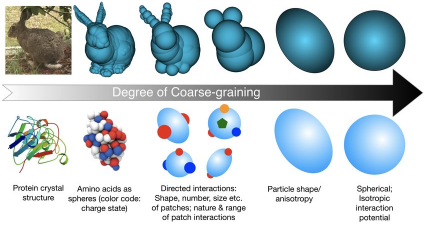
- Understand those *coarse-graining* maps preserving important relations, [4].
- Discover the conditional dependence among latents factors z_i, [5].



Learning Causal Representations

A priori, we must face the problem of what **representation** preserves the causal structure hidden on observations $\mathbf{X}$.

- Understand those *coarse-graining* maps preserving important relations, [4].
- Discover the conditional dependence among latents factors z_i, [5].



Implications

- **Semi-Supervised Learning:** $X \to Y$, SCM: $P(X, Y) = P(X)P(Y|X)$ but there is no information of $P(Y|X)$ from $P(X)$. In the anti-causal direction there can be information:

$$Y \to X, \quad P(X) \not\perp\!\!\!\perp P(Y|X)$$

- **Robustness and strong generalization:** learning autonomous modules to aid generalization out of distribution $P(X, Y) \to P^{\dagger}(X, Y)$, this is important for *strategic behaviour*.

but also for **Causal Discovery**, **Reinforcement Learning**, **Continual Learning** and **Scientific Applications**, [1].

Implications

- **Semi-Supervised Learning:** $X \to Y$, SCM: $P(X, Y) = P(X)P(Y|X)$ but there is no information of $P(Y|X)$ from $P(X)$. In the anti-causal direction there can be information:

$$Y \to X, \quad P(X) \not\perp\!\!\!\perp P(Y|X)$$

- **Robustness and strong generalization:** learning autonomous modules to aid generalization out of distribution $P(X, Y) \to P^{\dagger}(X, Y)$, this is important for *strategic behaviour*.

but also for **Causal Discovery**, **Reinforcement Learning**, **Continual Learning** and **Scientific Applications**, [1].

Implications

- **Semi-Supervised Learning:** $X \to Y$, SCM: $P(X, Y) = P(X)P(Y|X)$ but there is no information of $P(Y|X)$ from $P(X)$. In the anti-causal direction there can be information:

$$Y \to X, \quad P(X) \not\perp\!\!\!\perp P(Y|X)$$

- **Robustness and strong generalization:** learning autonomous modules to aid generalization out of distribution $P(X, Y) \to P^{\dagger}(X, Y)$, this is important for *strategic behaviour*.

but also for **Causal Discovery**, **Reinforcement Learning**, **Continual Learning** and **Scientific Applications**, [1].

Implications

- **Semi-Supervised Learning:** $X \to Y$, SCM: $P(X, Y) = P(X)P(Y|X)$ but there is no information of $P(Y|X)$ from $P(X)$. In the anti-causal direction there can be information:

$$Y \to X, \quad P(X) \not\perp\!\!\!\perp P(Y|X)$$

- **Robustness and strong generalization:** learning autonomous modules to aid generalization out of distribution $P(X, Y) \to P^{\dagger}(X, Y)$, this is important for *strategic behaviour*.

but also for **Causal Discovery**, **Reinforcement Learning**, **Continual Learning** and **Scientific Applications**, [1].

Conclusion

"Why can't we just train a huge model that learns environments' dynamics including all possible interventions? After all, distributed representations can generalize to unseen examples and if we train over a large number of interventions we may expect that a big neural network will generalize across them"

1. Causality offers an important complement: learn structures of data.
2. Generalization is tied to model's assumptions: in causal setting they become more explicit.

Conclusion

"Why can't we just train a huge model that learns environments' dynamics including all possible interventions? After all, distributed representations can generalize to unseen examples and if we train over a large number of interventions we may expect that a big neural network will generalize across them"

1. Causality offers an important complement: learn structures of data.
2. Generalization is tied to model's assumptions: in causal setting they become more explicit.

Conclusion

"Why can't we just train a huge model that learns environments' dynamics including all possible interventions? After all, distributed representations can generalize to unseen examples and if we train over a large number of interventions we may expect that a big neural network will generalize across them"

1. Causality offers an important complement: learn structures of data.
2. Generalization is tied to model's assumptions: in causal setting they become more explicit.

References


Towards Causal Representation Learning
Bernard Scholkopf *et al.* (2021), arXiv:2102.11107.


Structure by Architecture: Disentangled Representations without Regularization
Felix Leeb *et al.* (2021), arXiv:2006.07796.


CausalWorld: A Robotic Manipulation Benchmark for Causal Structure and Transfer Learning
Ossama Ahmed *et al.* (2020), arXiv:2010.04296.


Multi-Level Cause-Effect Systems.
Krzysztof Chalupka *et al.* (2015), arXiv:1512.07942.


Measuring Statistical Dependence with Hilbert-Schmidt Norms
Artur Gretton *et al.* (2005), Springer Berlin Heidelberg.

Thank you for listening!
and
Stay in contact with us!

Emanuele Marconato

emanuele.marconato@unitn.it
http://sml.disi.unitn.it/