A Category-theoretical Meta-analysis of Definitions of Disentanglement

Yivan Zhang[1, 2] Masashi Sugiyama[2, 1]

[1]The University of Tokyo

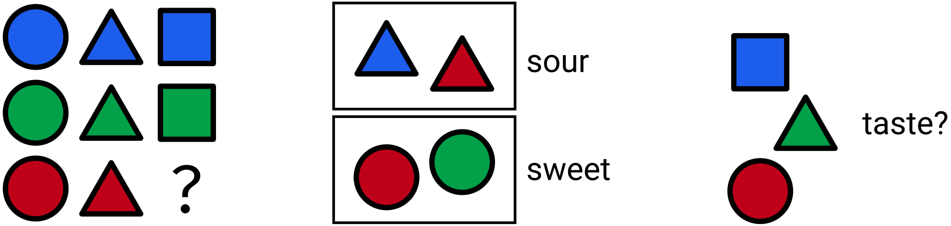
[2]RIKEN AIP

July 2023

International Conference on Machine Learning 2023

What is disentanglement?

Disentanglement: the process of identifying and separating the underlying factors of variation in data.



(a) Composition (b) Supervision (c) Generalization

- Factors: colors, shapes, and tastes
- Task: taste prediction based on colors and shapes
- We can predict unseen candies without observing all combinations.

Can a neural network do this?

Existing definitions

Algebraic approach: group theory, representation theory

Group actions capture the symmetries of an object [Cohen and Welling, 2014, 2015]. A disentangled encoder should be equivariant to group actions of a **direct product** of groups [Higgins et al., 2018].

Statistical approach: probability, statistics, information theory

Probabilistic models capture the relationships and uncertainty of variables. A disentangled encoder should satisfy certain statistical **independence** conditions [Higgins et al., 2017, Chen et al., 2018, Suter et al., 2019].

What do direct product of groups and independent random variables have in common?

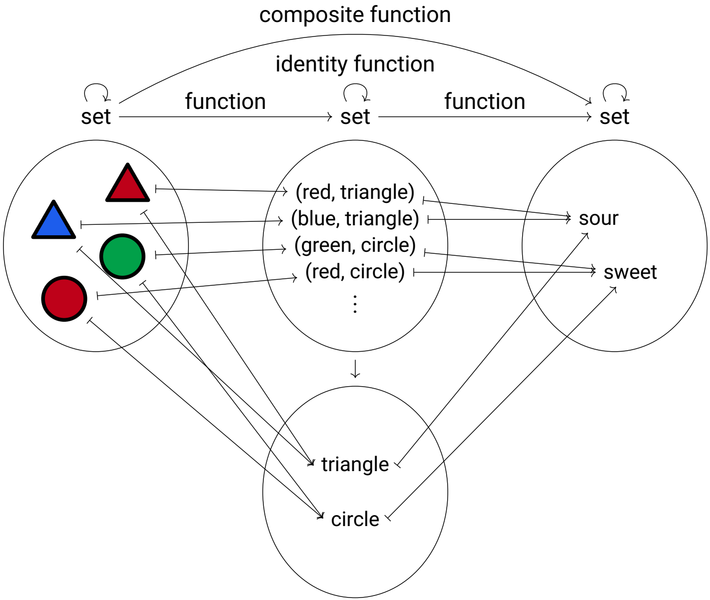
A unified perspective?

Questions

- What are the defining properties of disentanglement?
- Can we define disentanglement using only sets and functions?
- Are the existing algebraic and statistical approaches compatible?

Category theory: cartesian/monoidal product underlies many existing definitions of disentanglement.

Product: core of disentanglement



$\mathbf{Set}$: category of sets and functions

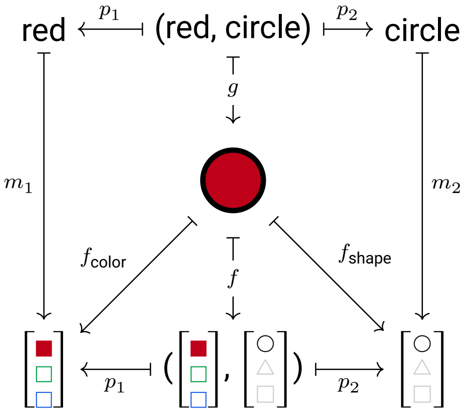
A function $C \to A \times B$
to a Cartesian product of sets
is just two component functions
$C \to A$ and $C \to B$.

A function $A \times B \to C$
from a Cartesian product of sets
can depend on both components.

When is $A \times B \to C$ just $A \to C$?

Defining properties of disentangled representations

Modularity: factor $Y \to$ code Z is a product morphism.



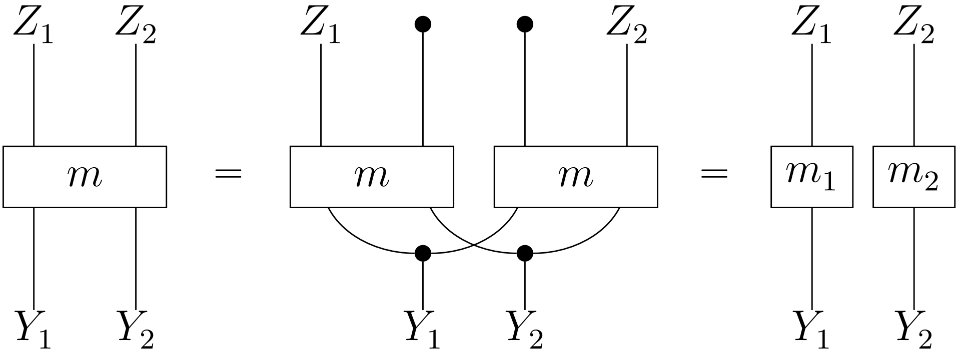
$$
\begin{tikzcd}
Y_1 \arrow[dd, "m_1"'] & Y_1 \times Y_2 \arrow[l, "p_1"'] \arrow[r, "p_2"] \arrow[d, "g"] & Y_2 \arrow[dd, "m_2"] \\
& X \arrow[dl, "f_1"'] \arrow[d, "f"] \arrow[dr, "f_2"] & \\
Z_1 & Z_1 \times Z_2 \arrow[l, "p_1"] \arrow[r, "p_2"'] & Z_2
\end{tikzcd}
$$

Informativeness: factor $Y \to$ code Z is a split monomorphism.

$$
\begin{tikzcd}
& {*} \circlearrowleft a \arrow[dddl, dashed, mapsto] \arrow[ddr, dashed, mapsto] & \\
& {*} \times {*} \circlearrowleft a \times b \arrow[u] \arrow[d] \arrow[ddl, dashed, mapsto] \arrow[ddr, dashed, mapsto] & \\
& {*} \circlearrowleft b \arrow[dl, dashed, mapsto] \arrow[ddr, dashed, mapsto] & Z_1 \circlearrowleft a_Z \\
X \arrow[loop above, "a_X"] \arrow[loop left, "(a\times b)_X"] \arrow[loop below, "b_X"] \arrow[urr, "\phi_*"] \arrow[rr, "f := \phi_{* \times *}"] \arrow[drr, "\phi_*"'] & & Z \circlearrowleft \begin{array}{l}(a\times b)_Z \\ = a_Z \times b_Z\end{array} \arrow[u] \arrow[d] \\
& & Z_2 \circlearrowleft b_Z
\end{tikzcd}
$$

algebra, action, and equivariance $\rightsquigarrow$ functors and natural transformations

Stochastic maps



measure, joint, and independence $\rightsquigarrow$ Markov category of stochastic maps

Conclusion

- Modularity, direct product, independence $\rightsquigarrow$ product in a category
- Formulation of disentanglement in more complex problems
- More structures and operations beyond product!

A Category-theoretical Meta-analysis of Definitions of Disentanglement

Yivan Zhang[1, 2] Masashi Sugiyama[2, 1]

[1]The University of Tokyo

[2]RIKEN AIP

July 2023

International Conference on Machine Learning 2023

References

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Neural Information Processing Systems*, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/1ee3dfcd8a0645a25a35977997223d22-Abstract.html. 2

Taco Cohen and Max Welling. Learning the irreducible representations of commutative lie groups. In *International Conference on Machine Learning*, 2014. URL https://proceedings.mlr.press/v32/cohen14.html. 2

Taco Cohen and Max Welling. Transformation properties of learned visual representations. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.7659. 2

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl. 2

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018. URL https://arxiv.org/abs/1812.02230. 2

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, 2019. URL http://proceedings.mlr.press/v97/suter19a.html. 2