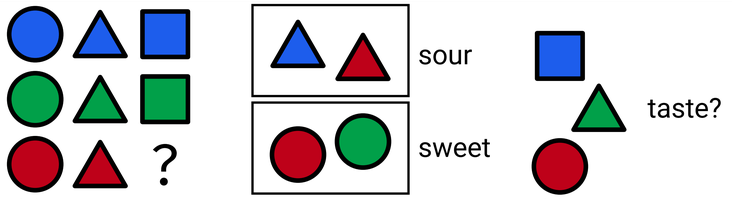
TITLE: A Category-theoretical Meta-analysis of Definitions of Disentanglement

TL;DR: # Disentanglement is a Product Morphism.

Yivan Zhang[1, 2] Masashi Sugiyama[2, 1]
[1]The University of Tokyo [2]RIKEN AIP

arXiv:2305.06886
https://yivan.xyz
yivanzhang@ms.k.u-tokyo.ac.jp

What is disentanglement?



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

- Colorful and tasty candies!
- We only need to taste a handful of candies to find out the relationship between their color, shape, and taste. We can use this knowledge to predict the taste of other candies.
- Can a neural network do this?

Algebraic definitions

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

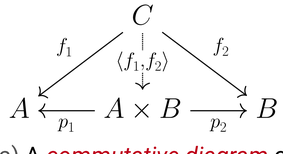
Statistical definitions

Probabilistic models capture the relations 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].

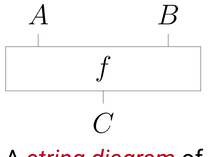
A unified definition?

- What do direct product and independent random variables have in common?
- Can we define disentanglement using only functions?
- What are the defining properties of disentanglement?

Category theory provides a suitable **abstraction** to identify, formalize, and organize common patterns, mathematically rigorous **diagrammatic reasoning**, and **generality** to tackle increasingly complex machine learning problems.



(a) A *commutative diagram* of a morphism $C \to A \times B$ to a *cartesian product*

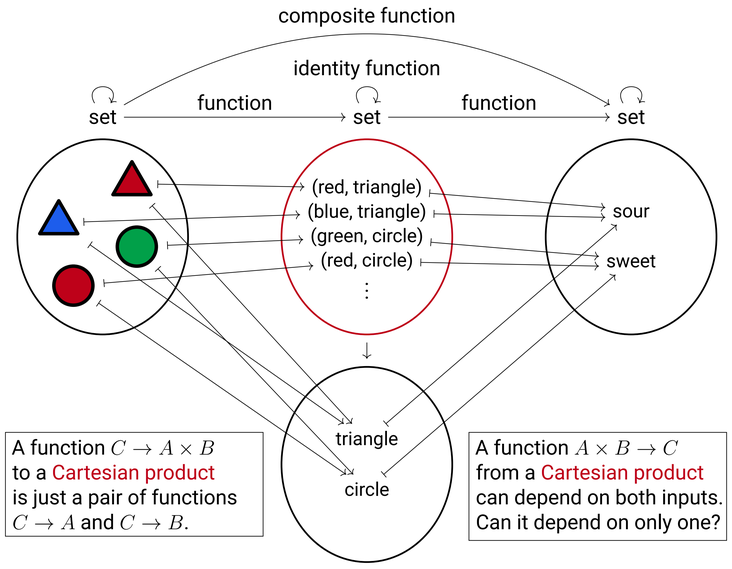


(b) A *string diagram* of a morphism $C \to A \otimes B$ to a *monoidal product*

Product: core of disentanglement

$\mathbf{Set}$: *category* of sets (*objects*) and functions (*morphisms*)
Let's consider Y: factors, X: observations, and Z: codes.

Disentanglement: $f : X \to Z$ is a morphism to a *product*.

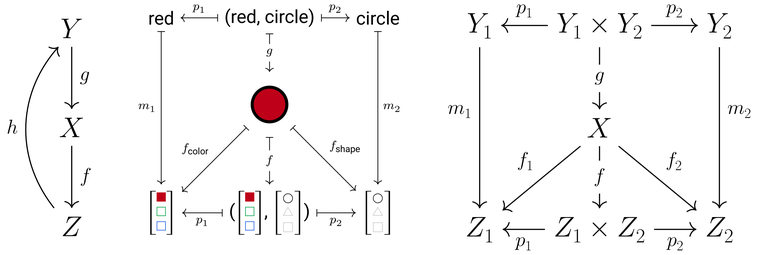


A function $C \to A \times B$ to a Cartesian product is just a pair of functions $C \to A$ and $C \to B$.

A function $A \times B \to C$ from a Cartesian product can depend on both inputs. Can it depend on only one?

Modularity: a code encodes only one factor

Modularity: $m : Y \to Z := f \circ g$ is a *product of morphisms*.



- When is $A \times B \to C$ just $A \to C$?
- We can use *exponential objects* and *pullbacks*.

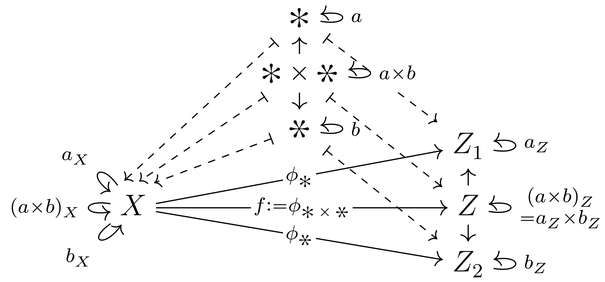
Informativeness: codes encode factors faithfully

Informativeness: $m : Y \to Z$ is a *split monomorphism*.

- $m : Y \to Z$ has a *retraction* $h : Z \to Y$, s.t. $h \circ m = \mathrm{id}_Y$.
- We should *disentangle* modularity and informativeness!

Equivariant maps

$[\mathbf{S}, \mathbf{Set}]$: category of functors and natural transformations

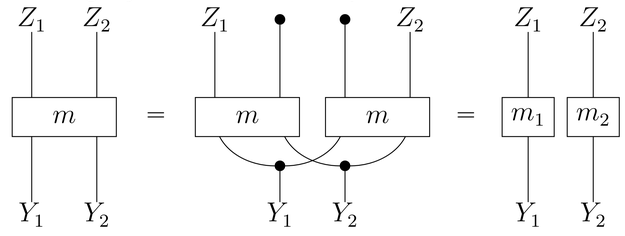


- Algebra action = *functor* from a *single-object category*
- Equivariant map = *natural transformation*

equivariance $\rightsquigarrow$ naturality

Stochastic maps

$\mathbf{Stoch}$: category of measurable spaces and stochastic maps



- Joint distributions are *monoidal products*, not cartesian.
- We can use copy & delete in a *Markov category* [Fritz, 2020].

probability & statistics $\rightsquigarrow$ Markov category

Next steps?

- Disentanglement metrics (*enriched category theory*?) [Zhang and Sugiyama, 2023]
- Analyses on functor categories and Markov categories
- More structures and operations beyond product!

References

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

Taco Cohen and Max Welling. Learning the irreducible representations of commutative lie groups. In *ICML*, 2014.

Tobias Fritz. A synthetic approach to markov kernels, conditional independence and theorems on sufficient statistics. *Advances in Mathematics*, 2020.

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 *ICLR*, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv:1812.02230*, 2018.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *ICML*, 2019.

Yivan Zhang and Masashi Sugiyama. Enriching disentanglement: Definitions to metrics. *arXiv:2305.11512*, 2023.