[Magic Stones] Normative Models Draft

Working Title: Theory formation in dynamic causal generalization

Bonan Zhao
b.zhao@ed.ac.uk

November 28, 2019

1 Models

Task formalization

A magic stone g is formalized as an expression $g = c_g \wedge s_g$, where its color c_g takes value from a finite color space $\mathbb{CL} = \{red, yellow, blue\}$, and its shape s_g takes value from a finite shape space $\mathbb{SP} = \{round, circle, triangle\}$. In total, the set of all magic stones G is of size $(|\mathbb{CL}| \times |\mathbb{SP}|) = 9$.

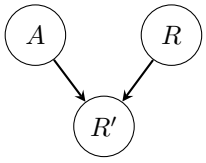

Figure 1: A magic stone task.

A magic stone task consists of an agent stone A that changes the recipient stone R into R'. An ordered set of (A, R, R') forms a data point in this model. There are $|G|^3 = 729$ outcomes in the entire sample space.

R' is conditional on its previous state R and the agent stone A, as shown in Figure 1 The joint distribution of $P(a, r, r') = P(r'|r)P(r)P(r'|a)P(a)$[1].

Let's use θ to refer to the parameterization representing agent stone A's causal power. Hence, the task to infer agent stone's causal power is to infer parameter θ given observed data points.

Causal power

Causal power θ is given by a function $f(A, R) = R'$, stating that given an agent stone A and recipient R, R changes into R'. In other words,

$$P(a, r, r'|\theta) = \begin{cases} 1 & \text{if } f_\theta(a, r) = r' \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

To better capture the uncertainty of real life, we use a softmax over Equation (1) to smooth out the cutoffs. Given A, R, there are 9 possible R's, hence set $b = e^9$ as the base for this specific softmax, yielding

$$P(a, r, r'|\theta) = \begin{cases} .992 & \text{if } f_\theta(a, r) = r' \\ .001 & \text{otherwise} \end{cases} \tag{2}$$

[1] Lazy notation P(a) stands for P(A=a).

Bayesian update

In a standard Bayesian context, causal power functions form the hypothesis space H. During the *learning* phase, the posterior distribution of causal power functions given data is calculated by:

$$P(h|D) = \frac{P(D|h)P(h)}{\sum_{h_i \in H} P(D|h_i)P(h_i)} \tag{3}$$

While *generalizing*, the posterior predictive distribution is given by marginalizing the distribution of new data $\tilde{d}$ over the posterior distribution (Z is the normalizing factor):

$$P(\tilde{d}|D) = Z \sum_{h \in H} P(\tilde{d}|h)P(h|D) \tag{4}$$

Hierarchical Bayesian models

What causal power functions people may inspect forms the key part of this model. We follow a hierarchical Bayesian model (HBM) framework to tackle this problem.

The HBM we consider here consists of three levels of abstractions. As illustrated in Figure 2, at the highest level of abstraction is theory space T, and each theory $t \in T$ generates some causal power functions, forming theory t's hypothesis space H_t. Hypotheses generate data D, at the lowest level of abstraction in this framework.

Under this framework, theories are learned via:

$$P(t|D) \propto \sum_{h_t \in H_t} P(h_t|D)P(h_t|t)P(t) \tag{5}$$

And generalization to new data point $\tilde{d}$ with posterior belief about theories is

$$P(\tilde{d}|D) \propto \sum_{t \in T} \sum_{h_t \in H_t} P(\tilde{d}|h_t)P(h_t|t)P(t) \tag{6}$$

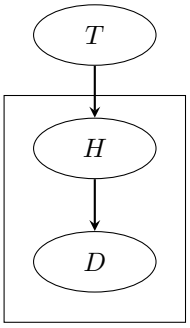

Figure 2: A HBM.

Candidate theories

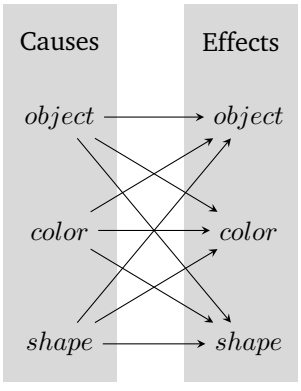

Figure 3: Formation space for f.

A causal power function deals with cause conditions, and defines what effects will take place. Both a specific object, or an abstract feature, can serve as a cause condition. Similarly, an effect is to change a stone into a specific state (another object), or change one of its features. (Changing both features is the same as changing to a specific object.)

Different arrow combinations in Figure 3 generates different sets of causal power functions. For example, take the arrow $object \rightarrow object$, it generates $|G|^2 = 81$ possible causal power functions. One of these functions is $f_1((red \wedge square), R) = (red \wedge square)$, meaning that if an agent stone is of color red and shape $square$, the recipient stone will turn into color red and shape $squre$.

One example of a feature-level arrow is $color \rightarrow shape$. This generates $|\mathbb{CL}| \times |\mathbb{SP}| = 9$ possible causal power functions. One of such functions is $f((red, s_A), R) = (c_{R'}, square)$, meaning that a red agent stone, regardless of its shape, changes the recipient stone to shape $square$ and keeps its original color.

Arrow	$o \to o$	$o \to c$	$o \to s$	$c \to o$	$c \to c$	$c \to s$	$s \to o$	$s \to c$	$s \to s$
$\|f\|$	81	27	27	27	9	9	27	9	9
Total									225

Table 1: Number of causal power functions each arrow combination produces. o: $object$, c: $color$, s:$shape$.

Conditions		Task 1	Task 1.1	Task 1.2	Task 2	Task 2.1	Task 2.2
Separate theories	$o \to o$	$f((r,t),R) = (r,s)$	(r,s)				
	$o \to c$	$f((r,t),R) = (r,R_s)$	(r,c)				
	$o \to s$	$f((r,t),R) = (R_c,s)$	(b,s)				
	$c \to o$	$f((r,A_s),R) = (r,s)$	(r,s)	(r,s)			
	$c \to c$	$f((r,A_s),R) = (r,R_s)$	(r,c)	(r,c)	$f((r,A_s),R) = (r,R_s)$	(r,c)	(r,c)
	$c \to s$	$f((r,A_s),R) = (R_c,s)$	(b,s)	(b,s)	$f((r,A_s),R) = (R_c,s)$	(b,s)	(b,s)
	$s \to o$	$f((A_c,t),R) = (r,s)$	(r,s)		$f((A_c,c),R) = (r,s)$	(r,s)	
	$s \to c$	$f((A_c,t),R) = (r,R_s)$	(r,c)		$f((A_c,c),R) = (r,R_s)$	(r,t)	
	$s \to s$	$f((A_c,t),R) = (R_c,s)$	(b,s)		$f((A_c,c),R) = (R_c,s)$	(b,s)	
Entire theory space							
Sub theory space?							

Table 2: Model predictions for sample experiment.

Table 1 shows the number of total causal power functions these arrows produce. We treat each of these arrow as a *theory*, and take all the possible causal power functions generated by all the theories as the complete hypothesis space H. Let $\mathbf{z}$ be a vector of 255 elements, for each $z_i \in \mathbf{z}$, $z_i = 1$ if $h_i \in H_t$, and $z_i = 0$ if $h_i \notin H_t$. The prior probability of a theory generates a hypothesis $P(h|t)$ is therefore defined as a softmax over $\mathbf{z}$:

$$P(h) = \sigma(\mathbf{z}) \tag{7}$$

Example

Experiment

Consider the following tasks

Learning task 1: $red\ triangle$ changes $yellow\ square$ to $red\ square$.

Generalization task 1.1: $red\ triangle$ changes $blue\ circle$ to ?

Generalization task 1.2: $red\ square$ changes $blue\ circle$ to ?

In addition to learning task 1:

Learning task 2: $red\ circle$ changes $yellow\ square$ to $red\ square$.

Generalization task 2.1: $red\ circle$ changes $blue\ triangle$ to ?

Generalization task 2.2: $red\ square$ changes $blue\ triangle$ to ?

Model predictions

As shown in Table 2. Lower case letters stand for r: red, b: blue, y: yellow, c: circle, s: square, t: triangle. Under 'Task n' columns are the winning hypothesis under each theory, and under 'Task $n.n$' columns are predicted data.

2 Limitations

1. Cannot account for non-observed changes.
2. Ignore recipient influence.

3 Alternatives

Code it in a neural net (input-output pairs) and compare the results.