Compositional Explanations: A Proposal

Joey Velez-Ginorio

June 26, 2018

1 Model

Given an arbitrary physical simulation, and an event we care about, a model of explanations should compute a distribution over explanations of that event. Something like the following list, but with probabilities associated with each item.

- *(Event) because (Cause1)*
- *(Event) because ((Cause1) or (Cause2))*
- *(Event) because ((Cause1) and (Cause2))*
- *(Event) because (((Cause1) and (Cause2)) or ((Cause3) and (Cause4)))*
- *....*

In order to get that distribution, we frame the task as a problem of bayesian inference. Formally, how do we compute:

$$P(Explanation|World, Event) \tag{1}$$

If we specify a generative model, we can use it to guide how we compute this distribution. With it, we also have a convenient way to partition the mechanisms of our model, separating parts that compute different elements of the distribution. This will be discussed after introducing the generative model in figure 1.

You can view each node in the generative model as a random variable, and each edge as encoding the statistical dependencies of those random

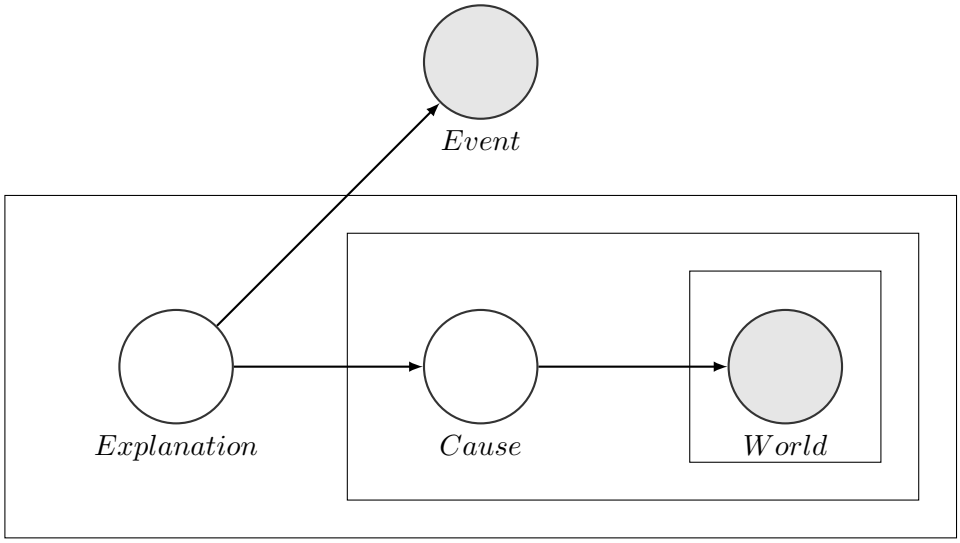

Figure 1: Our generative model

variables. We should be able to infer information about upstream nodes from observations of downstream nodes. Knowledge of the world lets us infer cause, knowledge of cause and the event of interest allow us to infer explanation. The boxes are meant to highlight that the generative model is composed of smaller models. At the core, is a model of the world provided by a physics engine, succeeded by a counterfactual-based model of cause, and lastly the model of explanation we pursue in this proposal.

The relationship between these random variables are captured nicely by using the generative model to derive a joint distribution. Note, from here on out, I will use abbreviated notation for the random variables: Ex for $Explanation$, C for $Cause$, Ev for $Event$, and W for $World$.

$$P(Ex, C, Ev, W) = P(W|C)P(C|Ex)P(Ev|Ex)P(Ex) \tag{2}$$

1.1 Posterior

With the joint distribution, we can return to our definition of equation 1, and define a posterior distribution over explanations:

$$
\begin{aligned}
P(Ex|W, Ev) &\propto P(W, Ev|Ex)P(Ex) \\
&\propto P(W|Ex)P(Ev|Ex)P(Ex) \\
&\propto \sum_C (P(W|C)P(C|Ex))P(Ev|Ex)P(Ex)
\end{aligned} \tag{3}
$$

Our definition of the posterior definition helps us capture two intuitions about explanations. The likelihood, $P(W, Ev|Ex)$, should favor explanations that invoke relevant causes and that mention the event we're interested in. The prior, $P(Ex)$, should favor simple over complex explanations. In the following sections we discuss what it means to invoke a relevant cause, and what it means for an explanation to be simple or complex – and how to include these intuitions in our computation.

1.2 Likelihood, $P(W, Ev|Ex)$

The likelihood should favor explanations that invoke relevant causes from what happened in the world, and that include the event we're interested in explaining. The expansion of the likelihood helps us achieve this. We will discuss each of the terms in more detail.

$$
P(W, Ev|Ex) = P(W|Ex)P(Ev|Ex) \tag{4}
$$

1.2.1 $P(Ev|Ex)$

The chance that an event occurred given an explanation can be treated as binary choice. It's 1 if the event was mentioned in the explanation and 0 if the event was not mentioned in the explanation.

1.2.2 $P(W|Ex)$

The chance that you saw what happened in the world because of causes invoked in an explanation would be high if the causes mattered. However, $P(W|Ex)$ does not express this intuition until we marginalize over all invoked causes in the explanation.

$$
P(W|Ex) = \sum_C P(W|C)P(C|Ex) \tag{5}
$$

Drawing on the approach of Gerstenberg et al. [1], we say that the probability of the world given a cause, $P(W|C)$, is high when the cause invoked is a *difference maker*. Otherwise, if the cause invoked would have made no or little difference to the observed world, it's value should be representative of the degree of influence. See [1] for specific details on what constitutes a cause as a *difference maker*. $P(C|Ex)$, can be treated as uniform for all causes, this term captures the assumption that all causes invoked in an explanation are equally important (not always true but reasonable in our task setting).

1.3 Prior, $P(Ex)$

The prior should favor explanations that are simple. To accomplish this we can set the prior to be negative exponential in the number of causes invoked in the explanation, x.

$$P(Ex) = e^{-x} \tag{6}$$

References

[1] T. Gerstenberg, N. D. Goodman, D. A. Lagnado, and J. B. Tenenbaum. How, whether, why: Causal judgments as counterfactual contrasts. In *CogSci*, 2015.