NORELLIEXPLANATORYLEARNINGEMPIRICISM

18 May 2025

Extracted Annotations (20/04/2022, 07:14:49)

"We formulate the challenge of creating a machine that masters a language as the problem of learning an interpreter from a collection of examples in the form (explanation; observations). The only assumption we make is this dual structure of data; explanations are free strings, and are not required to fit any formal grammar. This results in the Explanatory Learning (EL) framework described in Sec. 2" (Norelli et al 2022:2)

"Critical Rationalist Networks (CRNs), a family of models designed according to the epistemological philosophy pushed forward by Popper (1935). Although a CRN is implemented using two neural networks, the working hypothesis of such a model does not coincide with the adjustable network parameters, but rather with a language proposition that can only be accepted or refused in toto. We will present" (Norelli et al 2022:2)

"problem, where finite automata take the role of explanations, while regular sets are the phenomena. More recently, CLEVR (Johnson et al., 2017) posed a communication problem in a universe of images of simple solids, where explanations are textual and read like "There is a sphere with the same size as the metal cube". Another recent example is CLIP (Radford et al., 2021), where 400,000,000 captioned internet images are arranged in a communication problem to train an interpreter, thereby elevating captions to the status of explanations rather than treating them as simple labels3 . With EL, we aim to offer a unified perspective on these works, making explicit the core problem of learning an interpreter purely from observations." (Norelli et al 2022:3)

"many, the concept of explanation may sound close to the concept of program; similarly, the scientist problem may seem a rephrasing of the fundamental problem of Inductive Logic Programming (ILP) (Shapiro, 1981) or Program Synthesis (PS) (Balog et al., 2017). This is not the case. ILP has the analogous goal of producing a hypothesis from positive/negative examples accompanied by" (Norelli et al 2022:3)

"Such a model would assume that all the information needed to solve the task is embedded in the data, ignoring the explanations; we may call it a "radical empiricist" approach (Pearl, 2021). A variant that includes the explanations in the pipeline can be done by adding a textual head to the network. This way, we expect performance to improve because predicting the explanation string can aid the classification task. As we show in the experiments, the latter approach (called "conscious empiricist") indeed improves upon the former; yet, it treats the explanations as mere data, nothing more than mute strings to match, in a Chinese room fashion (Searle, 1980; Bender & Koller, 2020)" (Norelli et al 2022:6)