CS6360: Advanced Topics in Machine Learning

Group Theory $\cap$ Machine Learning

Pranav K Nayak

IIT Hyderabad

Paper Presentation 1

Homomorphism Autoencoder: Learning Group Structured Representations from Observed Transitions: Hamza Keurti, Hsiao-Ru Pan, Michael Besserve, Benjamin Grewe, Bernhard Schälkopf, *ICML 2023*

The paper attempts to model the effect of interventions as transformations in representation space.
They assert that this problem can be formulated as a problem of learning a homomorphism between the interventional structure of the world and the model's representations of it. This should allow it to be able to reverse-engineer the effects of potential interventions (transformations) through the knowledge of how its representations change.

The Learning Problem

- W is the latent space from which observations are generated, through the process g.
- O is the space of observations.
- Z is the space of representations, mapped to from O through the *inference process* h.

$$
\begin{tikzcd}
W \arrow[r, "{(G, \cdot_W)}"] \arrow[d, "b"] \arrow[dd, bend right, "f"'] & W \arrow[d, "b"] \arrow[dd, bend left, "f"] \\
O \arrow[d, "h"] & O \arrow[d, "h"] \\
Z \arrow[r, "{(G, \cdot_Z)}"] & Z
\end{tikzcd}
$$

Figure: The proposed group structure of the learning problem

HAE and Two Losses

Definition ($N-$step Prediction Loss)

$\mathcal{L}^N_{\text{pred}}(\rho, h) =$
$\sum_t \sum_{j=1}^N ||h(o_{t+j}) - (\prod_{i=0}^{j-1} \rho(g_{t+i}))h(o_t)||$

Definition ($N-$step Reconstruction Loss)

$\mathcal{L}^N_{\text{rec}}(\rho, h, d) =$
$\sum_t \sum_{j=1}^N ||o_{t+j} - d(\prod_{i=0}^{j-1} \rho(g_{t+i}))h(o_t)||$

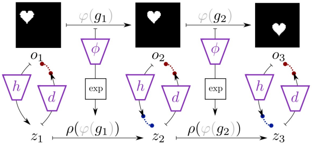

Figure: HAE's Architecture

A weighted sum of both losses, $\mathcal{L}(\rho, h, d) = \mathcal{L}_{\text{rec}^N}(\rho, h, d) + \gamma\mathcal{L}^N_{\text{pred}}(\rho, h)$, is optimized for.

Restrictions on the Representation Class

Theorem

If (ρ, h, d) are continuous and minimize the expectation of $\mathcal{L}^2_{pred}(\rho, h) + \gamma \mathcal{L}^k_{rec}(\rho, h, d)$, for $k \geq 0$, then ρ is a non-trivial group representation and (ρ, h) is a symmetry-based representation.

Informally, a symmetry-based representation is one that satisfies the following:

$\rho(g_1, g_2, ..., g_n)(z_1 \oplus ... \oplus z_n) = \rho_1(g_1)(z_1) \oplus ... \oplus \rho_n(g_n)(z_n)$ where $z_i = h(o_i)$.