On the Principles of Parsimony and Self-Consistency for the Emergence of Intelligence

Xueqing Yu

July 2022

1 context and motivation

world model 应既能从过去的经验中学习,也能解释当前环境的新输入。大脑的 world model 在结构和功能上都是高度模块化的 (structured anatomically and functionally)。

现有机器学习模型的弊端：高度同源的结构，暴力的训练方法；lack of richness in final learned representations;lack of stability in training(mode collapse);lack of adaptiveness and susceptibility to catastrophic forgetting;lack of robustness.

两个基本原则对应两个问题：

- parsimony–what to learn: information/coding theory
- self-consistency–how to learn: control/game theory

2 Two Principles for Intelligence

2.1 the principle of parsimony

The objective of learning for an intelligent system is to identify low-dimensional structures in observations of the external world and reorganize them in the most compact and structured way.

找一个变换 f，将高维（由多个非线性的低维子流形构成?）的输入映射到线性的低维流形子空间中：

- compression: 降维
- linearization: 将分布在非线性子流形上的 object 映射到线性子空间上
- sparsification: 稀疏化，将不同的类映射到具有独立或最大不连贯基的子空间（让不同子空间互相正交?）

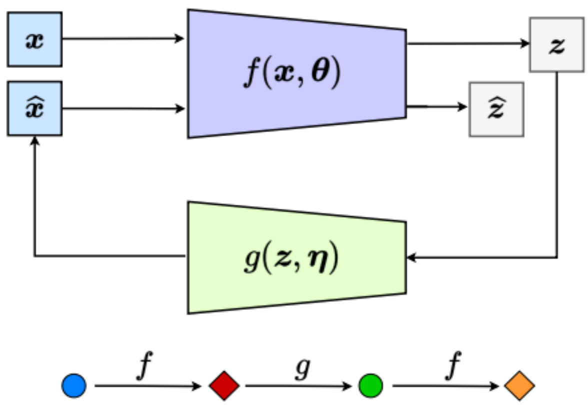
这种模型被称为 linear discriminative representation(LDR).
formulate: rate distortion
特征分布 $Z = [z^1, z^2, ...z^n]$,sampled data $X = [x^1, x^2, ...x^n]$, k 个类别，R^C 是 k 个类别 rate distortion 的平均，$R^C(Z) = \frac{1}{k}[R(Z_1) + ...R(Z_k)]$，$Z = Z_{12} \cup ..._k]$，则 rate reduction 定义为 $\Delta R(Z) = R(Z) - R^C(Z)$
通常情况下 R(Z) 难以计算,当样本从高斯分布取样时,有以下形式：$R(Z) = \frac{1}{2}logdet(I+\alpha ZZ^*)$，于是可以用 projected gradient ascent(PGA) 的方式求解

ReduNet,ReduNeXt 和上述更新方式有类似的结构，改写一下表达式可以发现 transformer 的 self-attention 实际上是在更新 rate distortion

2.2 the principle of self-consistency

An autonomous intelligent system seeks a most self-consistent model for observations of the external world by minimizing the internal discrepancy between the observed and the regenerated.

self-consistency 和 parsimony 两个原则必须是同时应用，若只是要减小 observed 和 regenerated 的差距，用参数多的模型很容易实现（过拟合）。



$f(x, \theta)$ 将 x 映射到特征子空间，$g(z, \eta)$ 将 z 映射回原始空间，如图形成一个闭环的反馈系统。通过比较 z 和 $\hat{z}$ 来评价重新生成的 $\hat{x}$ 和 x 的差距。z 和 $\hat{z}$ 的距离用 rate reduction 来衡量：

$$\Delta R(Z(\theta), \hat{Z}(\theta, \eta)) = R(Z \cup \hat{Z}) - \frac{1}{2}(R(Z) + R(\hat{Z}))$$

f 的目标：最大化 $\Delta R(Z)$，尽可能分辨 x 和 $\hat{x}$ 的差别，即最大化 $\Delta R(Z, \hat{Z})$, g 的目标：最小化 $\Delta R(Z, \hat{Z})$, 使得 regenerated 的数据和原始数据尽可能像，同时最小化达到目标所需的 $\Delta R(\hat{Z})$(why?)

于是，f 和 g 之间可以看作一种零和博弈，即：

$$\max_{\theta} \min_{\eta} \Delta R(Z) + \Delta R(\hat{Z}) + \Delta R(Z, \hat{Z})$$

(1)