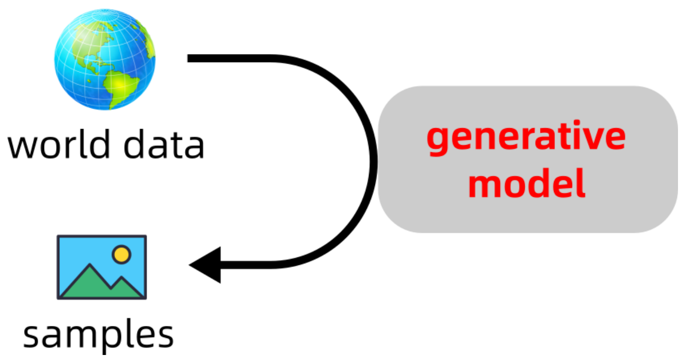
①

AGI = RL + LLM

Quick overview of generative models

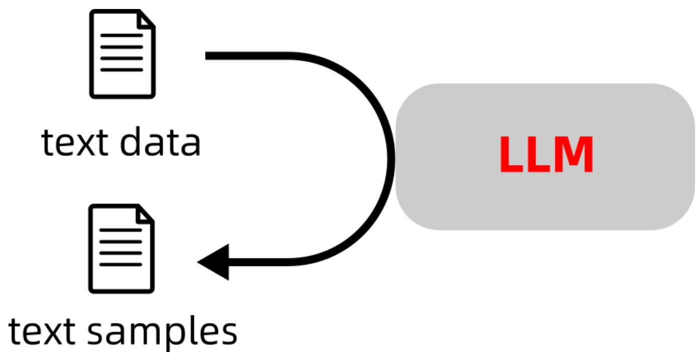
(This section is based from the book "The Science of Deep Learning" [Iddo Drori, 2023] with my own simplifications)

A generative model, as opposed to a classifying model, is one that learns the probability distribution of the data and outputs samplings from the learned distribution:



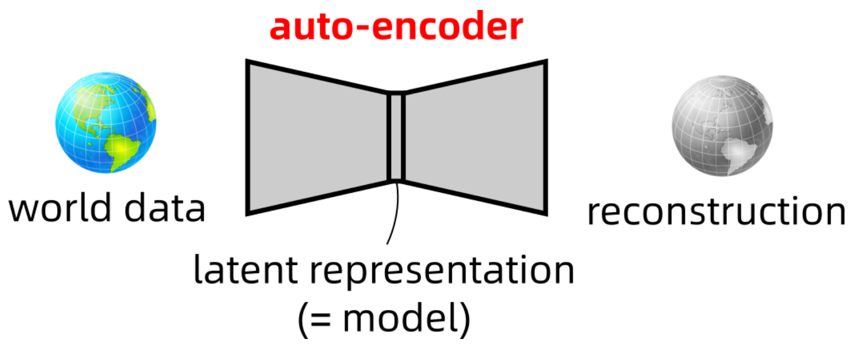
(1)

LLMs are a special case of generative models:



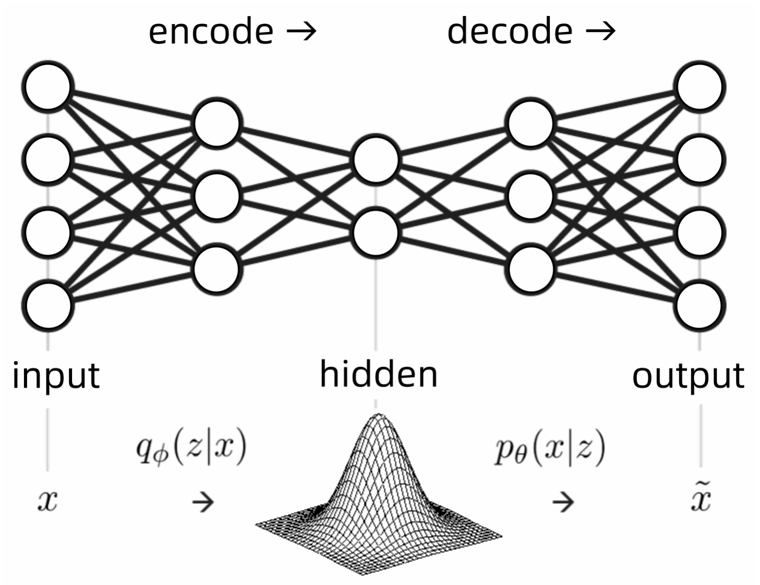
(2)

One class of generative models are auto-encoders, which forces information to flow through a narrow bottleneck, thus compressing the data into a compact, latent representation:



(3)

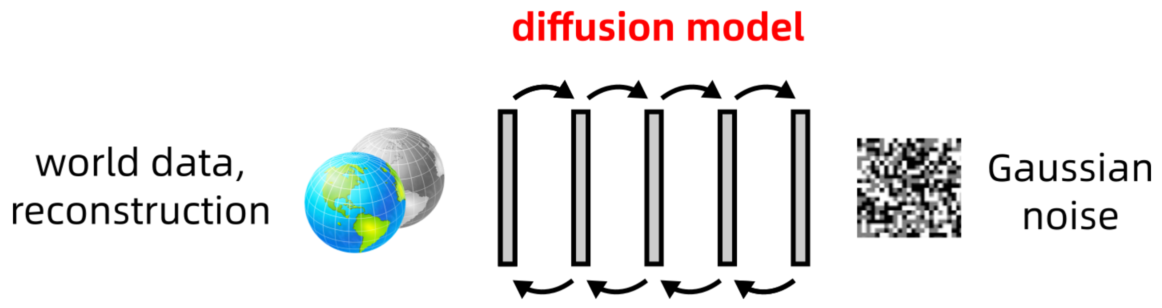
Of which, the VAE (variational auto-encoder) uses variational methods to find a probability distribution $q_\phi(z|x)$ that approximates the true 'posterior' distribution $q(z|x)$:



(4)

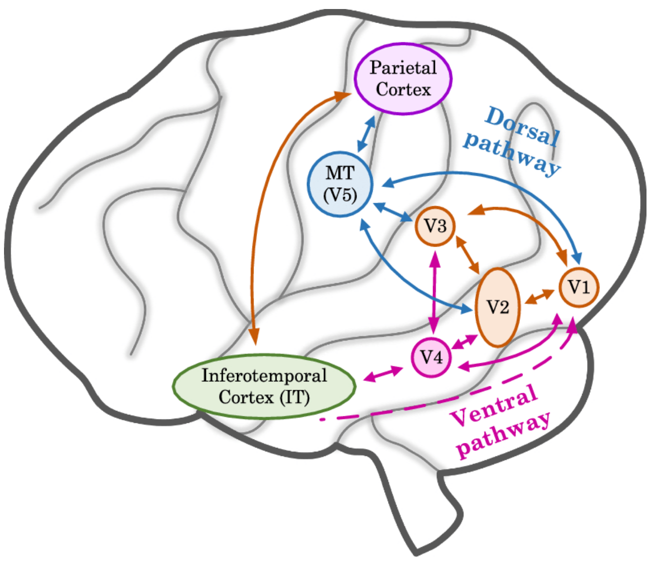
One variational inference algorithm recently proposed is SVGD (Stein variational gradient descent) which exploits efficiency in reproducing kernel Hilbert space.

Another generative model is the diffusion model, whose latent representation is distributed among its many layers:



(5)

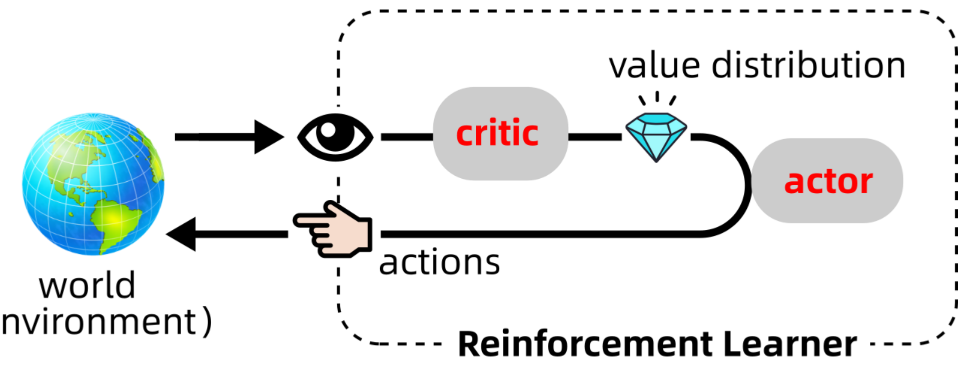
Interestingly, the human brain also has the structure of an auto-encoder. Sensory information is processed by a hierarchy of cortical areas, getting increasingly abstract representations, which information is then back-projected towards the primary sensory areas, thus forming an auto-encoder similar to (3) but folded in the middle.



(6)

Reinforcement learning

AGI should be developed under the framework of RL (reinforcement learning), which tries to find an optimal policy that acts in an environment, that maximizes the total reward over a time horizon:



(7)

In its most general form, an RL algorithm tries to maximize the following Bellman objective:

$$\max_{\pi} \mathop{\mathbb{E}}_{\substack{a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim p(\cdot|s_t,a_t)}} \left[\sum_t \gamma^t R(s_t, a_t) \right] \tag{8}$$

RL only needs to learn two things (probability distributions): the policy π which is concerned with "values", and the world model p which is concerned with "truths".

State-of-the-art RL algorithms tend to have an actor-critic structure, that simultaneously learns value functions (denoted Q or V) and policy functions (denoted π).

I tend to favor the SAC (soft actor-critic) algorithm for AGI because it has an elegant theoretical underpinning based on entropy maximization. If an RL algorithm always chooses the highest reward and does not explore the environment, such a strategy may turn out inferior to one that has some kind of "curiosity". So we add an entropy term H to the objective (8) to make it "soft", so the agent tries to maximize rewards as well as make its behavior most "random":

$$\max_{\pi} \mathbb{E}_{\pi} \left[\sum_t \gamma^t R(s_t, a_t) + \alpha H(\pi(a_t|s_t)) \right] \tag{9}$$

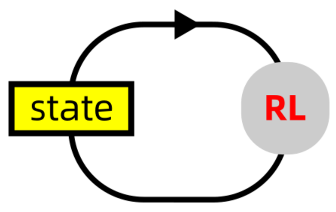
The function $Q(s,a)$ is a cross-section of the value function $V(s)$, which explicitly shows the choice of actions. If we always choose $\arg\max_a Q(s,a)$, such a strategy is purely exploitative. So we make Q into a probability distribution via Boltzmann's construction:

$$\pi(\mathbf{a}|\mathbf{s}) \propto \frac{\exp Q(\mathbf{s},\mathbf{a})}{Z} \tag{10}$$

where Z is the partition function. It turns out amazingly that this policy maximizes the objective (9)

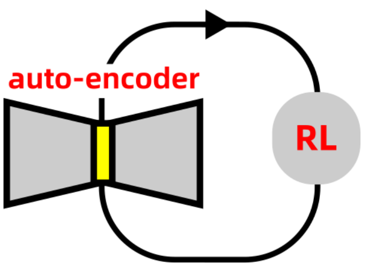
RL + World Model

The simplest RL architecture is like this:



(11)

I propose to integrate it with an auto-encoder like this:



(12)

which means the two algorithms shares the same state.

For example, if the auto-encoder sees an apple then it produces a high-level representation of apple in its hidden layer(s). If the RL desires to eat an apple it will produce the actions necessary to grab it.

One tricky issue here deserves explanation: I posited the RL to work in "mental space", whose actions are "thoughts" (but could also include real physical actions). So the RL not only learns how to act in the world, but also learns how to "think" based on rewards. After some thinking I concluded that this setup will not create inconsistencies.

Now the state (yellow box) is altered by two different algorithms with distinct objectives. After some considerations I also tend to think it is OK....

About our group

We operate as a DAO (decentralized autonomous organization) based on transparent operations and reward system based on weighted voting, to enable global collaboration without racial (or other forms of) discrimination.

We value: democracy, freedom of speech, racial equality, transparency, tolerance of mistakes, and a learning environment.

It is OK for anyone to challenge other member's theories, ideas, proposals, etc.