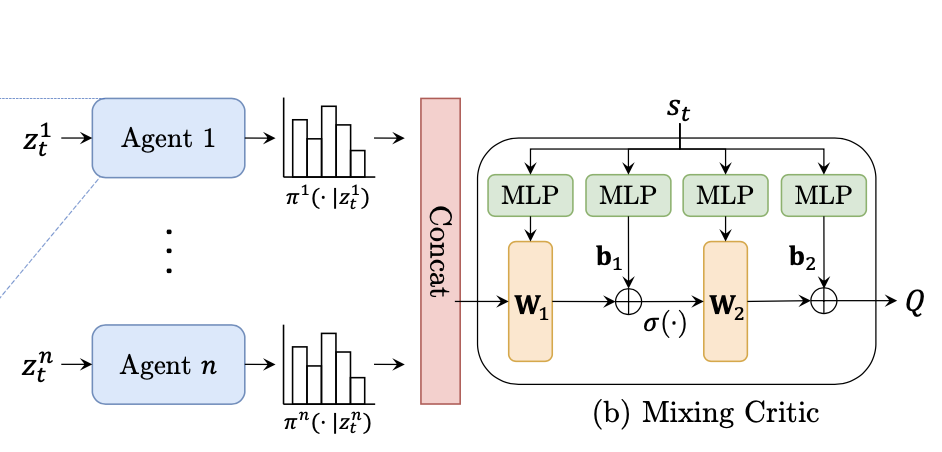
本周工作：

本周加入了nips那篇工作的两个改进，一个是把全局状态信息s与每个智能体的策略pi分开，通过下图的方式实现，而之前的做法都是把s和每个智能体的策略pi给直接串联起来，这么改进的原因是为了更好的利用全局状态信息。还有一个改进是自适应entropy。在加入这两个改进之后，跑了一个3天的实验后，发现已经可以跑出baseline的效果。在他们的代码中，我还看到了另一个改进便是gumble-softmax action selector，但是在论文中没阐述。调研后发现gumble-softmax主要是为了使得采样可导化。他们并没有利用这一点，因为其实他们的网络并不是端到端的。



下周工作：

目前已经基本复现了baseline的效果，我现在有一个想法是在他们的工作中，他们虽然写的是优化q值，其实是直接在优化v值，那么我们可以借鉴advantage = q - v 这种单智能体强化学习中常用的优化方法去减小方差，提升效果。同时加入gumble-softmax看一下效果。