Glance and Focus Networks for Dynamic Visual Recognition

·Q1：请解释输入一张图片后该网络的推理过程

输入一张图片后，首先resize得到一张缩略图，输入全局编码器提取全局特征，经过全局池化和分类器之后得到预测结果。

若此时的预测分数高于阈值，则以此时的预测结果为最终结果；否则基于全局编码器提取的特征，利用决策网络在原图中裁剪一小块区域。

将裁剪的小块儿区域输入局部编码器提取特征，经过全局池化和分类器之后得到预测结果。

若此时的预测分数高于阈值，则以此时的预测结果为最终结果；否则基于此时局部编码器提取得到的特征，利用决策网络在原图中重新裁剪一小块区域并重新分类，直到预测分数超过阈值或者决策网络的决策次数超过最大的裁剪次数。

·Q2：请解释IMPROVED TECHNIQUES FOR GFNET中的Contrastive Reward 是在解决什么问题

理想情况下，我们希望奖励函数的每一步决策都能使得奖励分数增量最大，但是实际情况中发现，在一些场景下，随机选择下一步决策的裁剪区域同样也会得到很大的奖励分数增量，这无疑会为决策网络的训练带来不确定性干扰。所以，把奖励函数改写为t+1时刻决策网络裁剪区域的预测分数和随机裁剪相对应的预测分数的差值，也就是希望决策网络的每一步决策相比于随机裁剪策略能得到明显的奖励分数增量，从而区分那些虽然能带来很大的奖励增量但却是随机裁剪的策略。