**本周主要工作：**

1. **SSAST: Self-Supervised Audio Spectrogram Transformer实验**
2. **Contrastive Learning based Hybrid Networks for Long-Tailed Image Classification实验**
3. **Unet++网络学习及伯努利采样和neighbor down-sampler**
4. **Mamba网络和Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model**

**1.SSAST: Self-Supervised Audio Spectrogram Transformer**

该论文以音频和语音分类为研究重点，利用未标记的数据集进行自监督学习（判别掩码和恢复掩码），网络结构如下：

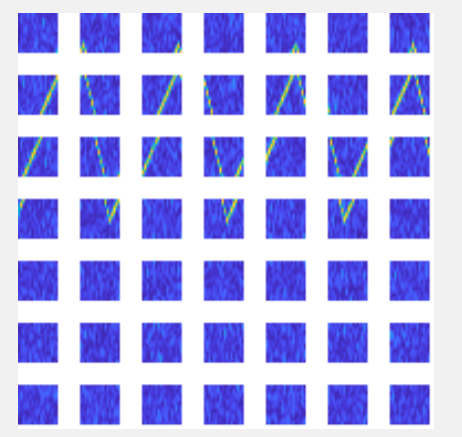


实验设置：

在预训练阶段，batchsize设置为256，学习率设置为0.1，图像块大小为16\*16，图像输入为112\*112，因此一张图片被分成49块，随机掩盖其中30块进行分类预测（论文中图像输入为1024\*128，按照16\*16分割，共有512块，通过实验证明选取250-400结果较好，及50%-80%）。Transformer编码层使用Vit transformer

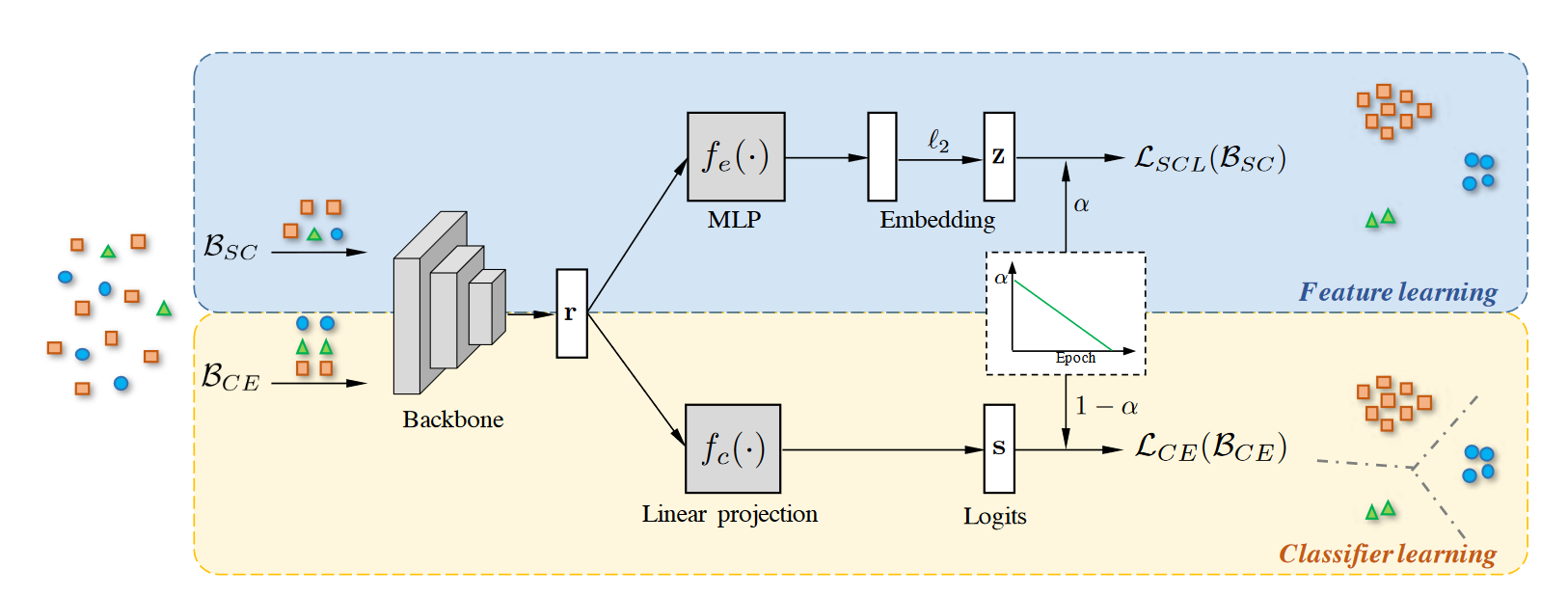
实验结果：

预训练准确率为3%。



**2. Contrastive Learning based Hybrid Networks for Long-Tailed Image Classification**

该论文创新点在于将网络学习分为特征学习（监督对比损失学习）和分类器学习（交叉熵损失学习），网络结构如下所示：



实验设置：

主体网络采用RestNet18，batchsize为128，训练次数为1000轮，设置，初始学习率为0.01。

实验结果：

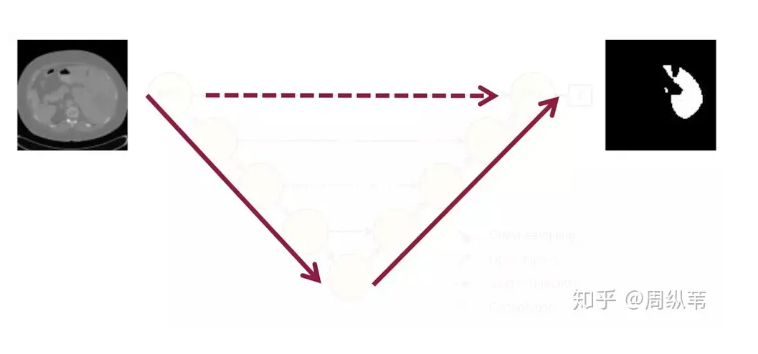
在训练到44轮时，准确率能达到80%，但会提示显存不足，如果则不会出现这个问题。

Loss = epoch/n\_epochs \* loss1 + (1-epoch/n\_epochs) \* loss2

**3. Unet++网络学习及伯努利采样和neighbor down-sampler**

**unet++网络结构**

先介绍unet网络，主要由编码器（下采样）、解码器（上采样）和skip connection构成，其中编解码结构在最开始主要是用来图像压缩和去噪，目的在于通过储存一个特征和解码器实现存储图像。



**为什么要用下采样？/下采样有什么用**

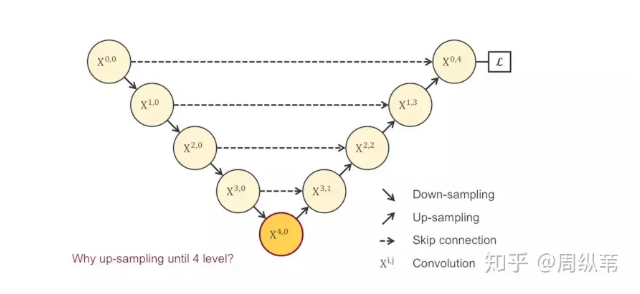
下采样是缩小图像，用更少的信息表示原图像。下采样的方法包括对区域内信息取最大值，取均值，随机取，加权取。我认为下采样的作用主要为增大感受野（建立远像素间的相关性），减少计算量（图像特征变少了）。

**为什么要用上采样？**

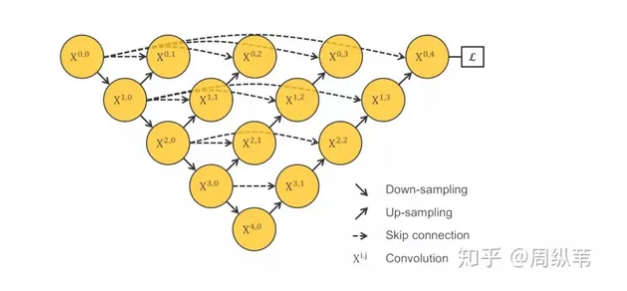
上采样利用特征得到一个想要的输出，可以说是恢复图像。

**skip connection的作用？**

将浅层特征与输入进行拼接在送往下一层，为了避免梯度消失。另一种避免梯度消失的方法是残差结构（adding），本质上都是找到一条捷径使得梯度可以直接反传到输入上。



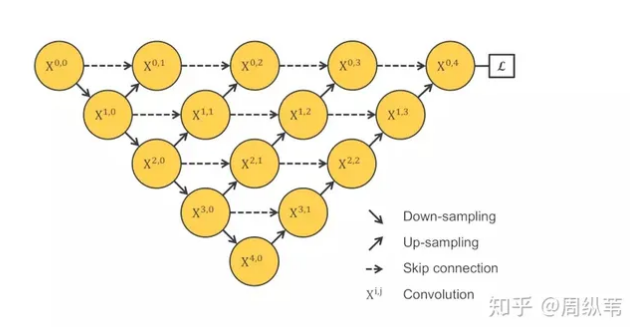
**为什么降采样是四层，有没有什么依据？在图像分割中，深层特征和浅层特征同样重要，有没有其他的方式将两者融合在一起**



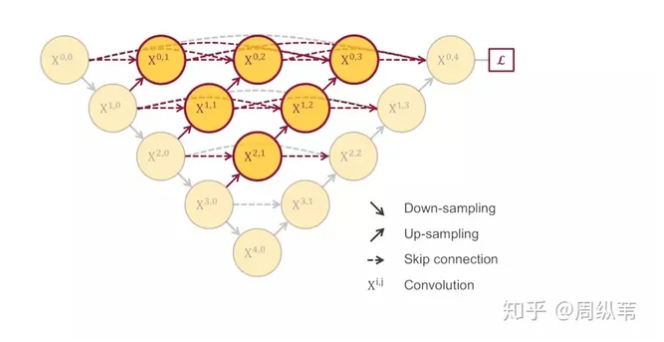
**这个网络中间部分是没有办法训练的，怎么修改？**

一种是加入deep supervision(深监督)

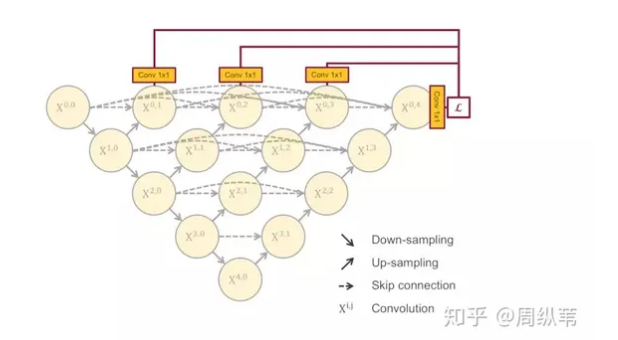
另一种是像下面这样连接，从而让梯度进行回传



进而，认为网络的长连接是有用的



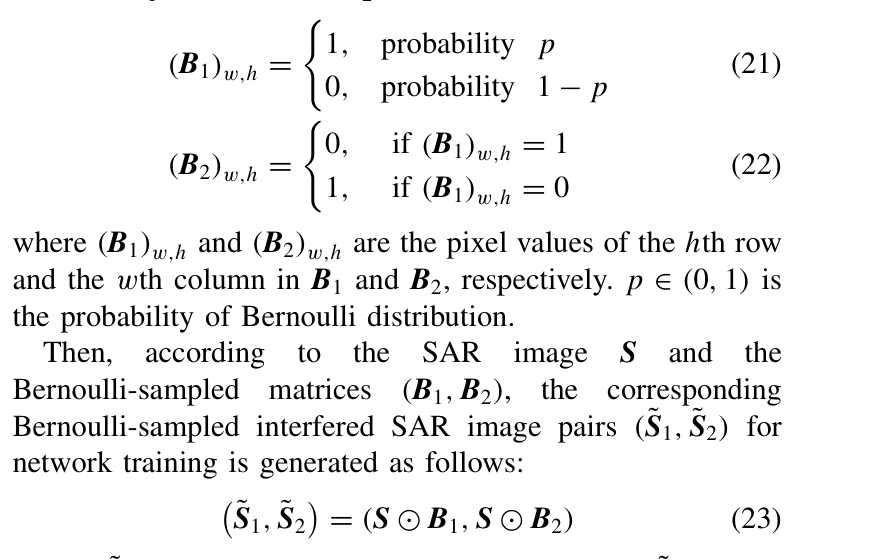
刚刚还提到了deep supervision,应用后就是这样



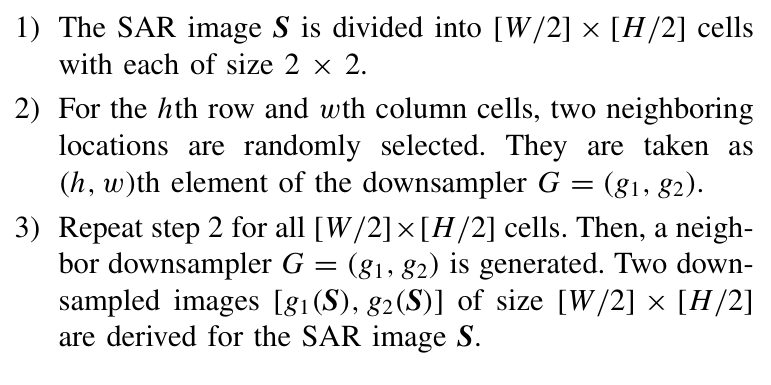
这个的好处就是可以剪枝

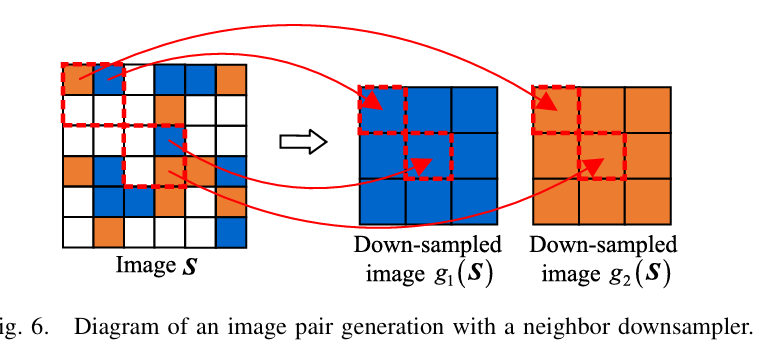
**伯努利采样和Neighbor down-sampler**，可以用在自监督学习中，用来学习图像相邻像素信息。

伯努利采样核心在于构建一个图片大小的伯努利分布，B1:（H,W）大小，每个像素0或者1。将B1与源图像的乘积作为s1，（1-B1）与原图像乘积作为S2，目的是通过s1使得网络得到s2.



Neighbor down-sampler是将图像分割成（H/2,W/2）个图像块，每个图像块中选取相邻的两个像素分别作为g1和g2，原理与上文相似，通过g1得到g2.







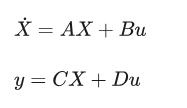
**4. Mamba网络和Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model**

**什么是Mamba？**

Mamba是有选择处理信息+硬件感知算法+更简单的SSM架构（状态空间模型）

**什么是状态空间模型？**

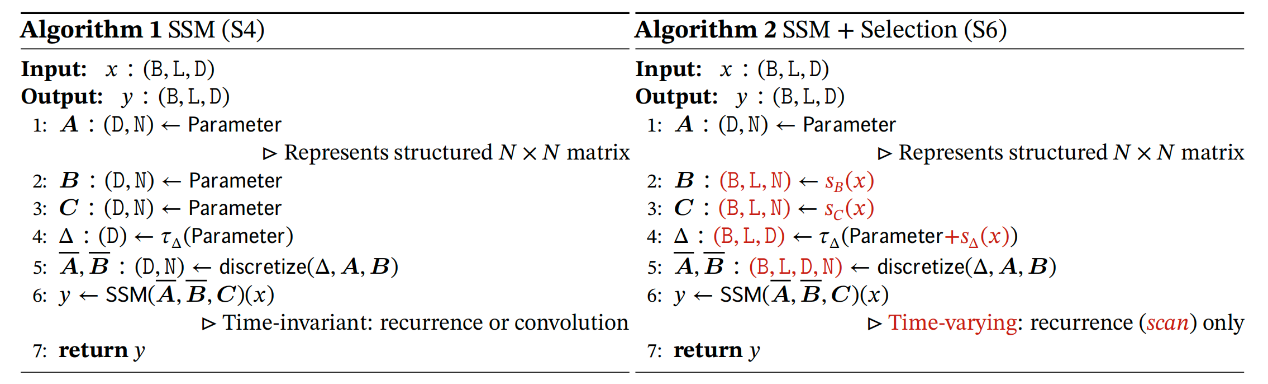
状态空间模型是一种对系统的动态描述，假设通过物理手段观测到量测向量，状态向量是描述系统动态特征的向量，两者结合在一起，才能对系统的动态特征作完整而充分的描述。反映了系统内部状态，而且揭示了系统内部状态与外部输出和输入变量的联系。



其中X表示状态，u表示输入，也就是说模型的输出量、状态量都可以由系统当前状态和输入量计算出来

**有选择处理信息是怎么做的？**

作者将状态空间方程的参数（B,C等）变成与输入有关的参数，之前（S4）是与输入无关的参数。

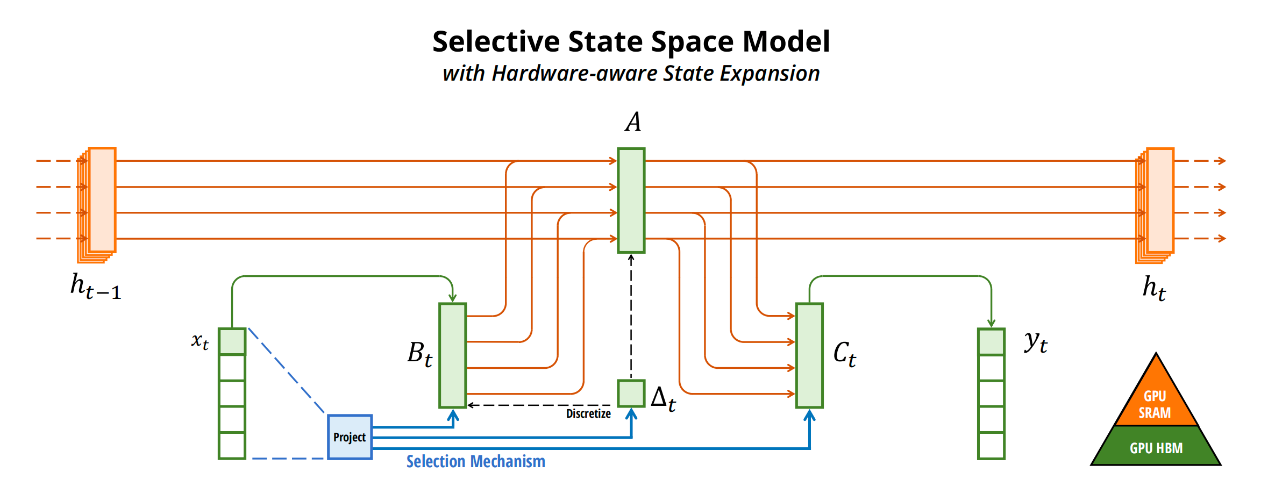


其中S6参数计算方式如下，虽然A不是直接用输入计算，但在与B更新中也会与输入有关。

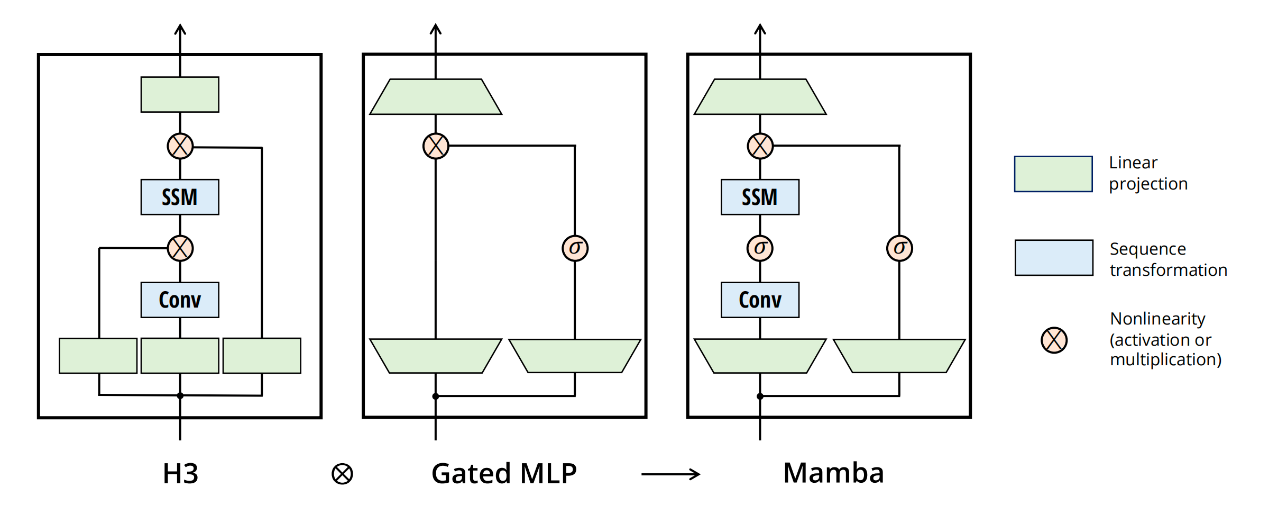

**硬件感知的扩展**

利用内存不同层级结构处理SSM的状态，减少高带宽但慢速的HBM内存反复读写的瓶颈。也就是把SSM参数计算放在快速SRAM中，计算完后在放到HBM中；通过并行扫描实现并行化；中间状态（SSM参数）不保存，每次重新计算。



**简化的SSM结构**

将大多数SSM架构比如H3的模块和transformer中的门控 MLP结构在一起，组成Mamba块，重复这个块，与归一化和残差连接结合，便构成了Mamba架构。mamba右边是skip connection。



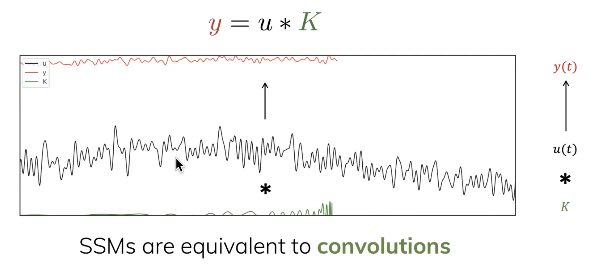
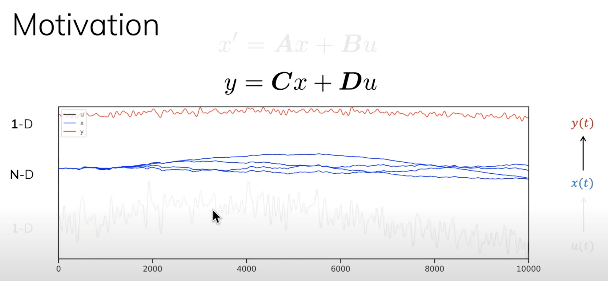
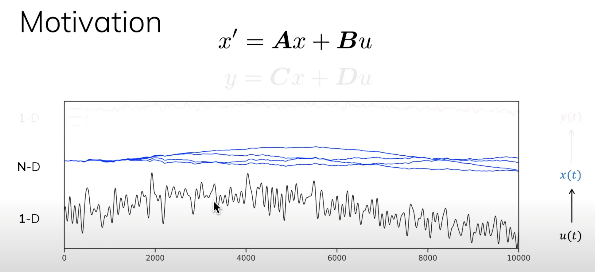
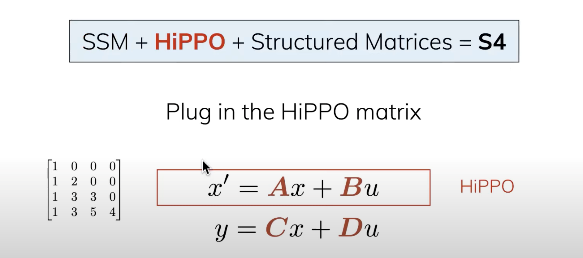
mamba下面的两个线性连接层用来将外部输入维度转换为内部输入维度，卷积用来将标志相互融合，否则它们将在每个维度上独立运行状态空间模型

**Mamba是如何实现上下文信息提取的？**

让模型最信息选择性处理，可以关注或忽略特定的内容，即使状态大小固定也能压缩上下文信息。可以认为Mamba每次参考前面所有内容的概括，越往后，舍弃细节，保留大意。

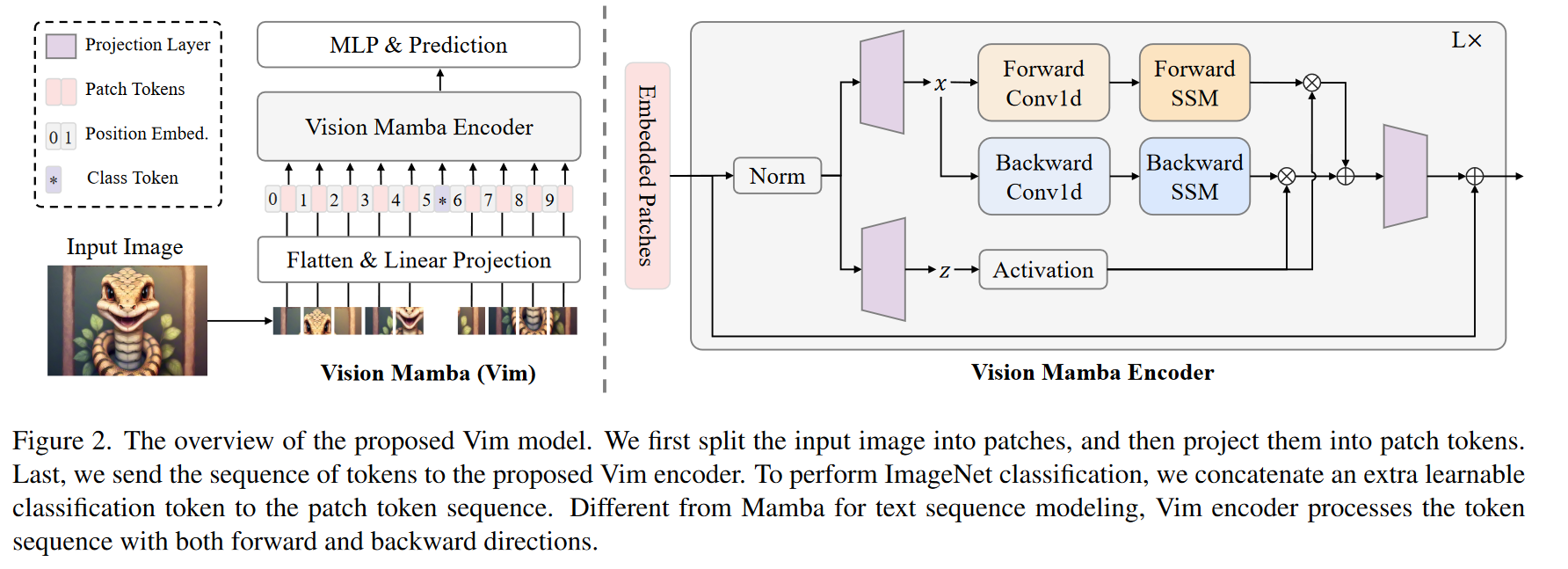
**为什么SSM能够提取上下文信息？**

对于低阶连续信号，可以用多项式去拟合输入，像HiPPO就是这样的思想，将输入映射到一个一个函数中。对于高阶信号，如果通过增加多项式的维度会引起维度爆炸，所以引入S4，将输入映射到状态函数之后，再由状态函数映射到输出。并且S4还通过skip connection将输入直接映射到输出。

**Vision Mamba**

解决视觉数据由于位置敏感性和视觉理解对上下文信息的要求导致SSM表示带来的困难。核心思想：1）为了解决视觉数据的位置敏感性，使用VIT的操作，将图片切块后添加位置编码。2）针对视觉上下文信息，**使用双向mamba模型**，一个从头开始，一个从尾开始，这样使得输出可以考虑前面信息也包含后面信息。



**为什么采用双向mamba获取上下文信息？**

因为如果只使用单个mamba的话，当前图像块序列的输出至于之前的图像块有关，跟之后的无关，所以为了弥补图像下文信息，参考双向RNN，设计了双向mamba模型，使得输出不仅与前向输入有关，还与后向输入有关。

