在深度学习（尤其是序列建模，如 Transformer、RNN）中，**掩码张量**（Mask Tensor） 是一种用于 “屏蔽” 或 “忽略” 输入序列中特定元素的技术，通过设置掩码值（如 0 或 -∞）告诉模型哪些位置的信息无效，应被忽略。其核心作用是控制模型对输入数据的关注范围，避免无效信息干扰计算。

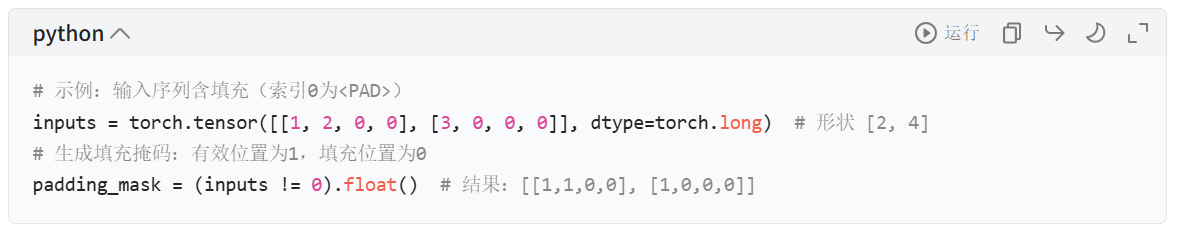
常见类型及用途

**1. 填充掩码（Padding Mask）**

场景：处理长度不一致的序列时，需用<PAD>符号补全短序列至相同长度（如文本句子、时序数据）。

作用：屏蔽填充符号（<PAD>）对应的位置，避免模型关注这些无意义的填充值。

实现：掩码张量形状与输入序列一致（如 [batch\_size, seq\_len]），填充位置标记为False（或0），有效位置标记为True（或1），在计算时过滤掉False位置的信息。



**2. 序列掩码（Sequence Mask / Look-ahead Mask）**

场景：在自回归任务（如语言生成）中，模型预测下一个元素时，不应看到未来的信息（如生成第i个词时，不能利用i+1及以后的词）。

作用：**屏蔽序列中 “当前位置之后的元素”，确保模型仅基于历史信息预测**。

实现：生成一个下三角矩阵（对角线及以下为1，以上为0），形状为 [seq\_len, seq\_len]，确保每个位置只能关注自身及之前的元素。



**3. 注意力掩码（Attention Mask）**

场景：在注意力机制（如自注意力）中，需同时过滤填充符号和未来信息。

作用：结合填充掩码和序列掩码，在计算注意力权重时，让无效位置的权重为 0（或接近 -∞，经 softmax 后变为 0）。

实现：通常将掩码转换为 -1e9 等极小值，与注意力分数相加，经softmax 后无效位置的权重被压制为 0。



**核心作用总结**

* 过滤无效信息：忽略填充符号（Padding Mask），避免无意义数据干扰模型。
* 控制时序依赖：在生成任务中屏蔽未来信息（Sequence Mask），保证预测的合理性。
* 优化注意力计算：通过掩码让模型聚焦于有效位置，提升计算效率和准确性。

**应用场景**

机器翻译、文本生成（需屏蔽未来信息）；

批处理不同长度的序列（需屏蔽填充）；

对话系统、推荐系统（需过滤无关元素）。

掩码张量是序列建模中不可或缺的技术，其设计直接影响模型对序列结构的理解能力。

np.tril()下三角矩阵； np.triu()上三角矩阵

