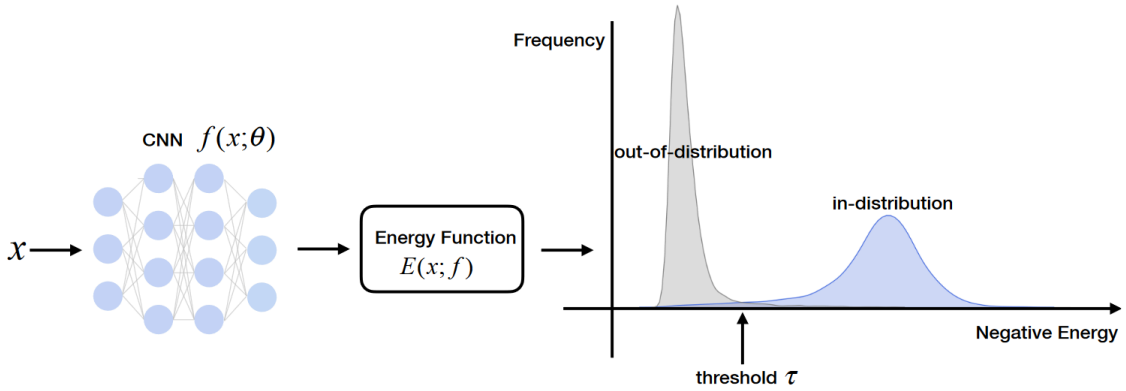
首先什么是能量分数方法介绍：

一个模型区分OOD和ID数据的主要依据是它们的能量分数，这个就是energy-based方法。

大致流程是这样：



首先有一个x，然后通过分类网络（通常是CNN）训练之后得到了对各个分类的概率结果（向量，其中每个logits对应着一个类别的可认为是概率的数值）。接着这个分类结果会喂进能量函数进行打分，最后对每个输入的结果按照阈值划分为ID 和OOD

这里看下能量分数是啥。

对于已经通过神经网络训练好的分类函数f，我们就有对每个输入的分类结果 $f(x)$ 以及对应的x
我们可以定义对应的能量分数计算为：

$$E(x;f) = -Tlog(\sum_{i=1}^{K} e^{f_i(x)/T})$$

其中K是对应的分类数量。T是温度参数，用来使得到的能量分数较为平滑。

从中可以看出ID数据的能量分数比较小，取负之后比较大。

这里有段证明为什么可以将能量分数作为区分OOD的依据。
我们训练模型损失函数选用NLL loss，根据能量分数的定义

$$\text{energy } E(\mathbf{x}, y) = -f_y(\mathbf{x}),$$

通过一些推导证明可以知道模型在最小化损失函数NLL的过程中也最小化了之前通过定义得到的能量分数：

$$E(\mathbf{x}; f) = -T \cdot \log \sum_{i}^{K} e^{f_i(\mathbf{x})/T}.$$

因此我们可以拿这个模型训练出来的结果喂给能量函数计算分数，然后以此为依据进行OOD检测。

并不是在所有情况下OOD和ID数据的能量分数都有显著差距的，为了使他们的差距拉大，我们还可以用有一定限制的能量分数方法，即强制规定将较低的分数分配给ID，较高的数据分配给OOD。

简单来讲在有条件限制下的能量分数方法我们的目标是找出一套参数θ 使得这个式子最小：

$$\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}\left[-\log F_y(\mathbf{x})\right] + \lambda \cdot L_{\text{energy}}$$

其中L_{energy}是与能量分数相关的正则化损失：

$$\begin{aligned} L_{\text{energy}} = {} & \mathbb{E}_{(\mathbf{x}_{\text{in}},y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}\left(\max(0, E(\mathbf{x}_{\text{in}}) - m_{\text{in}})\right)^2 \\ & + \mathbb{E}_{\mathbf{x}_{\text{out}}\sim\mathcal{D}_{\text{out}}^{\text{train}}}\left(\max(0, m_{\text{out}} - E(\mathbf{x}_{\text{out}}))\right)^2 \end{aligned}$$