# 神经网络最末隐藏层的信息：表示与校准应用

Luo-Ye Zou　Xin-Yu Hu
Nanjing University

## 摘要

神经网络隐藏层节点的输出往往被视作从输入提取的特征，并且最后几层的隐藏层被认为与输出层的输出之间共享了相当多的信息。我们注意到，当最末隐藏层激活函数使用 sigmoid 时，该层节点在各次前馈中的输出集中化明显，因此我们用二项分布与高斯混合分布独立编码其中每个节点的输出。我们计算了各节点编码结果与最终输出之间的互信息，以数值实验方式说明了特征表示的有效性。最后，我们利用编码得到的分布进行新任务的置信度估计，展示了最末隐藏层信息的校准潜力。

## 1. 引言

在神经科学中，对不同刺激下的神经元输出进行编码是一种常见的手段，根据编码得到的分布可以估计神经元对不同刺激的响应，以此推断神经元对某一任务的表示能力。相关工作中互信息被用以衡量神经元与任务之间的关联程度。在神经网络中，隐藏层节点的输出往往被视作从输入提取的特征。为分析隐藏层信息，任务方差 [4]（Task Variance）、掩膜法（mask）是一些常见手段，它们并不涉及输出层信息。而根据信息瓶颈（Information Bottleneck）[2] 理论，最后几层的隐藏层与输出层的输出之间共享了相当多的信息。我们推想，学习良好的神经网络之中，最末隐藏层单个节点的输出应该与标签有着显著的互信息。本文的工作中，我们针对激活函数使用 sigmoid 的最末隐藏层，用二项分布、高斯混合分布对其输出进行编码。我们计算了各节点编码结果与最终输出之间的互信息，数值实验表明互信息、任务方差与偏置之间有着显著的正相关，说明了特征表示的有效性。我们还将编码运用于校准，同样得到了不错的结果。

## 2. 最末隐藏层节点的信息表示

### 2.1. 编码

Sigmoid 是一种常见的激活函数，其输出在输入较大时会趋于饱和。因此在前馈过程中，经 sigmoid 激活后的节点输出容易表现出的集中化的分布。我们考虑两种神经网络，双隐层简单神经网络和 VGG16。它们分别处理 MNIST 数据集和 Cifar100 数据集的任务。我们让训练后模型对训练集进行预测，当模型给出了正确的结果，我们就记录下最末隐藏层的输出。双隐层简单神经网络的最末隐藏层宽度为 32，输出为 10 个类别，测试准确率为 97%。我们发现，在记录的隐藏层输出中，几乎大部分节点的输出值都在 0 和 1 附近。具体而言，大部分输出值集中在 0.9 到 1 或 0 到 0.1 之间，其他的输出值数量约占 3%。因此，我们用二项分布编码该神经网络的最末隐藏层节点输出，即将小于 0.5 的节点编码为 0，大于 0.5 的节点编码为 1。第二个神经网络 VGG16 是经典卷积神经网络。它的最末隐藏层宽度为 512，输出为 100 个类别，测试准确率为 74%。VGG16 的隐藏层输出没有集中在 0 和 1 附近，但依旧便于编码。我们用高斯混合模型对隐藏层每个节点编码，并测试了拟合的 R-square 误差。当对不同输出下的同一节点分别编码时，三个高斯分布的混合模型的平均拟合效果达到了 95.2%。当对不同输出下的同一节点统一编码时，十个高斯分布的混合模型的平均拟合效果达到了 81.1%。因此，我们用高斯混合模型编

码该神经网络的最末隐藏层节点输出。

### 2.2. 互信息、任务方差与偏置

信息瓶颈（Information Bottleneck）[2] 理论认为，在神经网络学习过程中，隐藏层与输入的互信息缩小，而隐藏层与标签的互信息增大。所以可以推想，经过良好训练的神经网络在前馈训练集数据时，最末隐藏层单个节点的输出也应该与标签有着显著的互信息。为衡量最末隐藏层单个节点的表示能力，我们计算各节点编码结果与最终输出之间的互信息:

$$I(h_i, y) = H(h_i) + H(y) - H(h_i, y) \tag{1}$$

其中 $h_i$ 为第 $i$ 个隐藏层节点的输出，$y$ 为标签，$H$ 为熵。对于高斯混合模型，

$$\begin{aligned} H(y) &= \int_y p(y)\log p(y)dy \\ &= \int_y (\sum_w w p_w(y))\log(\sum_w w p_w(y))dy \end{aligned}$$

不易计算。注意到

$$\begin{aligned} H' &= (\sum_w w p_w(y))\log(\sum_w w p_w(y)) \\ &\geq (\sum_w w p_w(y))\sum_w w\log p_w(y) \\ &= \sum_{w_1,w_2} w_1 w_2 p_{w_1}(y)\log p_{w_2}(y) \end{aligned}$$

其中 $w$ 为高斯混合模型的权重，$p_w(y)$ 为相应的概率密度函数。因此，我们用该下界来近似计算，得到

$$\begin{aligned} H(y) \approx \sum_{w_1,w_2} w_1 w_2 ( &\frac{(\mu_1-\mu_2)^2+\sigma_1^2}{2\sigma_2^2} + \\ &\frac{1}{2}\log\sigma_2 + \log 2\pi) \end{aligned} \tag{2}$$

其中 $\mu$ 为高斯分布均值，$\sigma$ 为标准差。$H(h_i, y)$ 则分解为 $H(y|h_i) + H(h_i)$，也做同样处理。任务方差（Task Variance）[4] 能有效表示神经网络节点的活跃度，定义为不同数据下隐藏层输出的变化幅度:

$$Var[h_i] = E[h_i^2] - E[h_i]^2 \tag{3}$$

偏置（Bias）是神经网络模型内部参数。最末隐藏层到输出层的函数可以写为

$$f(h) = Wh + b$$

其中 $W$ 为权重，$b$ 为偏置。为了记录此处偏置对隐藏层的影响，将上述函数改写为

$$f(h) = W(h + b') \tag{4}$$

使得偏置直接作用于隐藏层。求解不定方程 $b = Wb'$ 时，我们假定同一偏置对每个隐藏层节点的影响相同。实验数据为 MNIST 和 CIFAR100，分别使用双隐层神经网络和 VGG16 网络学习，计算其隐藏层各个节点的互信息、任务方差和偏置，再计算三者之间的皮尔逊相关系数。结果如表 2 所示，互信息、任务方差与偏置之间有着显著的正相关。可知在经过训练的神经网络最末隐藏层中，变化幅度大的节点会与输出共享更多信息，说明活跃的节点的确含有更多的特征信息。并且，由于偏置的高低可视为对节点的激励与抑制 [3]，可以认为神经网络模型学习到了这种模式。以下两表为隐藏层各个节点的互信息（MI）、任务方差（TV）和偏置（Bias）之间的皮尔逊相关系数。

| $\rho$ | MI | TV | Bias |
|---|---|---|---|
| MI | 1 | 0.949 | 0.493 |
| TV | 0.949 | 1 | 0.472 |
| Bias | 0.493 | 0.472 | 1 |

表 1. MNIST 数据集任务

| $\rho$ | MI | TV | Bias |
|---|---|---|---|
| MI | 1 | 0.674 | 0.057 |
| TV | 0.674 | 1 | 0.099 |
| Bias | 0.057 | 0.099 | 1 |

表 2. CIFAR100 数据集任务

## 3. 校准测试

容易想到，由于隐藏层的线性变换结果即为最终输出，因此利用隐藏层信息做贝叶斯估计的效果与利用最终输出估计的效果应该十分相近，或者说，利用隐藏层信息做贝叶斯估计就是在学习最后的线性层的参数。所以我们并不准备用最末隐藏层的输出来复现(replicate)信号检测理论(Signal Detection

Theory）。我们将以校准测试（Calibration Test）为例，展示最末隐藏层信息的表示能力。校准测试是指对于一个分类器，使其输出的概率分布与真实概率分布一致。过往的研究充分表明，神经网络的softmax 输出总是会高估预测概率，即表现得过于自信（overconfident）。我们猜测，学习良好的神经网络在面对常态的、与训练数据相近的输入时，隐藏层输出会有一定的模式（pattern），而面对异常的、分布外的输入时，隐藏层输出会有所不同。也就是说，神经网络隐藏层里含有能够表示神经网络置信度的信息。因此，我们用前述编码结果来估计测试数据的预测置信度。具体而言，我们定义神经网络成功学习的训练数据构成一个分布，符合这一分布的数据集合记为 $A$，我们假设某个新数据属于这一集合，即 $x \in A$，其输出结果为 $y$。计算其隐藏层输出的对数似然：

$$L(h|x \in A, y = l) = \sum_i L(h_i|x \in A, y = l) \quad (5)$$

作为新数据的预测置信度。同时，为了更好的拟合隐藏层输出模式，我们改进了用于编码的分布。共三种方案：三项分布（三类分别是小于 0.1、大于 0.9 和其他）、Beta 分布 $B(\alpha, 1-\alpha)$（$\alpha$ 为待估计参数）和高斯混合模型。我们对双隐层神经网络使用前两种方案，对 VGG16 网络使用高斯混合模型。我们用 AUROC（Area Under the Receiver Operating Curve）来衡量校准效果，基准线 [1] 为 Softmax 输出的置信度，AUROC 值越高，校准效果越好。实验数据为 MNIST 和 CIFAR100 的测试集，分别使用双隐层神经网络和 VGG16 网络学习。如下表所示，我们的方法在两个数据集上都取得了高于基准线的校准效果。

| 数据集 | Softmax 输出 | 三项分布似然 | Beta 分布似然 | 高斯混合模型似然 |
|---|---|---|---|---|
| MNIST | 0.953 | 0.960 | 0.958 | / |
| CIFAR100 | 0.864 | / | / | 0.873 |

表 3. 不同数据集上校准的 AUROC 指标

## 4. 总结

在本文中，我们深入研究了神经网络最末隐藏层的信息表示及其在校准任务中的应用。通过对双隐层简单神经网络和 VGG16 网络的分析，我们发现：

- 信息表示：最末隐藏层节点的输出在经过适当编码（如二项分布和高斯混合分布）后，与最终输出具有显著的互信息。这表明，隐藏层节点的输出有效地表示了输入数据的特征。
- 互信息、任务方差与偏置的关系：我们计算了隐藏层节点的互信息、任务方差和偏置之间的皮尔逊相关系数，结果表明它们之间存在显著的正相关关系。这说明活跃度高的节点往往与输出共享更多的信息，而偏置对节点的影响也能反映神经网络对特定模式的学习效果。
- 校准应用：我们进一步利用隐藏层的编码结果进行校准测试。实验结果表明，基于隐藏层输出的置信度估计能够在 MNIST 和 CIFAR100 数据集上获得比传统 Softmax 输出更好的校准效果。这验证了隐藏层信息在提高模型置信度估计准确性方面的潜力。

总之，我们的研究展示了神经网络最末隐藏层的节点输出在特征表示和校准任务中的重要作用。这一发现为进一步优化神经网络的结构和训练方法提供了新的思路，并为提高模型的置信度估计提供了有效途径。未来的研究可以探索更多类型的激活函数和编码方法，以进一步提升神经网络在复杂任务中的表现。

**致谢**. 感谢各位作者和老师的大力支持，以及使用的数据集提供者。

## 参考文献

[1] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. M. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. Zhu. A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56:1513–1589, 2021. 3

[2] N. Tishby and N. Zaslavsky. Deep learning and the information bottleneck principle. 2015 IEEE Information Theory Workshop (ITW), pages 1–5, 2015. 1, 2

[3] E. Williams, A. H.-W. Ryoo, T. Jiralerspong, A. Payeur, M. G. Perich, L. Mazzucato, and G. Lajoie.

Expressivity of neural networks with random weights and learned biases. ArXiv, 2024. 2

[4] G. R. Yang, M. R. Joglekar, H. F. Song, W. T. Newsome, and X.-J. Wang. Task representations in neural networks trained to perform many cognitive tasks. Nature Neuroscience, 22:297 – 306, 2019. 1, 2