- 方法0：原始Transformer
- 方法1：MLP处，第m层残差连接为$\frac{\sum_{i=1}^{m-1}\vec{y_i'}+\vec{y_m'}}{m}+\vec{y_m}$。Attention处残差不变
- 方法2：Attention处：第m层残差为$\frac{\sum_{i=1}^{m-1}\vec{x_i'}+\vec{x_m'}}{m}+\vec{x_m}$，MLP处残差不变。
- 方法3：微调方法1：将每一层权重$\frac{1}{m}$改为可学习参数并输出最后一层的参数列表。具体来说：除第0层外，每层都拥有一个独立的权重列表，第m层权重列表长度为m+1，涵盖了0-m层所有层的可学习权重。对权重做Softmax归一化后得到每层权重分布。

实验方法：

已验证方法0、1、2性能排序为：方法1>方法0>方法2。所以只对比方法2和方法3，并输出方法3的最终权重情况

实验数据

Method1

```
***** eval metrics *****
  epoch                   =        5.0
  eval_accuracy           =     0.5151
  eval_loss               =     2.4321
  eval_perplexity         =    11.3829
  eval_runtime            = 0:00:03.59
  eval_samples            =        143
  eval_samples_per_second =     39.773
  eval_steps_per_second   =      5.006
```

Method3

```
***** eval metrics *****
  epoch                   =        5.0
  eval_accuracy           =     0.5152
  eval_loss               =     2.4224
  eval_perplexity         =    11.2731
  eval_runtime            = 0:00:03.73
  eval_samples            =        143
  eval_samples_per_second =     38.293
  eval_steps_per_second   =       4.82
```

Conclusion:

方法3（方法1的微调版本）训练效果略好于方法1。从可学习权重的训练情况看，普遍规律是当前层权重最高，第0层次之，中间层很低。但仍存在例外情况（如Method3_2的第7层最高，第6层次之，其他层很低）

Paramerters_learned

Method3_v2 可学习权重参数保存文件

保存时间: 2025-07-11 15:39:04

模型配置信息:

- 总层数: 12
- 隐藏层大小: 768
- 注意力头数: 12
- 中间层大小: 2048

第 0 层 (索引: 0):
无权重参数 (第0层)

第 1 层 (索引: 1):
权重参数数量: 2
原始权重参数: [0.13120152056217194, 0.9348462224006653]
归一化权重: [0.3092464208602905, 0.6907535791397095]

第 2 层 (索引: 2):
权重参数数量: 3
原始权重参数: [0.3155142068862915, -0.6749102473258972, 1.1377571821212769]
归一化权重: [0.27419668436050415, 0.10184186697006226, 0.6239614486694336]

第 3 层 (索引: 3):
权重参数数量: 4
原始权重参数: [0.3958671987056732, -1.0396918058395386, -0.13419955968856812, 1.4354689121246338]
归一化权重: [0.21483947336673737, 0.051127996295690536, 0.12644712626934052, 0.6075854301452637]

第 4 层 (索引: 4):
权重参数数量: 5
原始权重参数: [0.6144229173660278, -0.9161863923072815, -0.8417204022407532, -0.16073720157146454, 1.6909453868865967]
归一化权重: [0.2064143568277359, 0.04466884955763817, 0.048122141510248184, 0.0950806513428688, 0.6057140827178955]

第 5 层 (索引: 5):
权重参数数量: 6
原始权重参数: [0.5879045128822327, -0.9200857281684875, -1.2028061151504517, -0.8659761548042297, 0.10349584370851517, 1.8702311515808105]
归一化权重: [0.17114679515361786, 0.037884101271629333, 0.0285544041544199, 0.03999046981334686, 0.10543692857027054, 0.6169872283935547]

第 6 层 (索引: 6):
权重参数数量: 7
原始权重参数: [0.6990680694580078, -0.8572046756744385, -1.343756914138794, -1.11174476146698, -0.6539031863212585, 0.05045885592699051, 1.8938844203948975]
归一化权重: [0.17894534766674042, 0.03774328902363777, 0.02320239320397377, 0.029261354357004166, 0.04625219851732254, 0.09354767203330994, 0.5910477042198181]

第 7 层 (索引: 7):
权重参数数量: 8
原始权重参数: [0.6566194891929626, -0.799759566783905, -1.4232546091079712, -1.0986571311950684, -0.757400631904602, -0.3214274048805237, 0.5068822503089905, 1.9425365924835205]
归一化权重: [0.1508520543575287, 0.03516039997339249, 0.018848348408937454, 0.026076197624206543, 0.03668174520134926, 0.05672713369131088, 0.12987367808818817, 0.5457803606987]

第 8 层 (索引: 8):
权重参数数量: 9
原始权重参数: [0.31510111689567566, -0.9053520560264587, -1.4948941469192505, -1.0270774364471436, -0.8735260367393494, -0.5996504426002502, -0.29848191142082214, 0.735927939414978, 2.028594493865967]
归一化权重: [0.09961925446987152, 0.029397282749414444, 0.016303183510899544, 0.026028104126453 4, 0.030347928404808044, 0.03990908712148666 4, 0.05393461138010025, 0.1517421007156372, 0.5527185201644897]

第 9 层 (索引: 9):
权重参数数量: 10
原始权重参数: [-0.08414673060178757, -0.9550889134407043, -1.4212254285812378, -0.9176802635192871, -0.7907712459564209, -0.6987264156341553, -0.6402156949043274, 0.21048875153064728, 0.7319492697715759, 1.9867534637451172]
归一化权重: [0.06553267687559128, 0.027429161593317986, 0.017209647223353386, 0.028474681079387665, 0.032327696681022644, 0.03544453904032707, 0.03758029639720917, 0.08798658847808838, 0.14821217954158783, 0.5198025703430176]

第 10 层 (索引: 10):
权重参数数量: 11
原始权重参数: [-1.3428415060043335, -1.5202676057815552, -1.7202295064926147, -1.1223571300506592, -1.0118205547332764, -0.92179936170578, -0.8545064926147461, -0.12556326389312744, 0.2906498312950134, 0.9937364459037781, 2.0450315475463867]
归一化权重: [0.017616892233490944, 0.014752792194485664, 0.012079021893441677, 0.021962638944387436, 0.02452957071363926, 0.026840191334486008, 0.028708504512906075, 0.05950954928994179, 0.09022890031337738, 0.1822603940963745, 0.5215115547180176]

第 11 层 (索引: 11):
权重参数数量: 12
原始权重参数: [-1.3069716691970825, -1.2758393287658691, -1.39922297000885, -0.9561449289321899, -0.832477331161499, -0.7357361316680908, -0.688753068447113, -0.24660183489322662, -0.07059293240308762, 0.19123457372188568, 0.5906198024749756, 1.7204997539520264]
归一化权重: [0.020957326516509056, 0.021620042622089386, 0.019110484048724174, 0.029764458537101746, 0.03368264436721802, 0.03710396960377693, 0.038888826966285706, 0.060513004660606384, 0.07215865701436996, 0.09375591576099396, 0.1397814303636551, 0.43266332149505615]