Method1

Attention处残差连接与原始模型相同。修改MLP处的残差为先前层重算的MLP输出累加。先前层MLP输出的重算方法为：

- 保留第一次Attention计算的attn_weights，W^V.weight， W^O$.weight，仅更换输入嵌入矩阵X
- 输入嵌入X做input_norm
- Attention重算
- Attention残差连接
- post_attn_layernorm
- MLP计算
- MLP计算结果不做残差连接直接输出，作为重算后的MLP输出

Method2

MLP处残差连接与原始模型相同。修改Attention处的残差为先前层重算的Attention输出累加。先前层Attention输出的重算方法为：

- 保留第一次Attention计算的attn_weights，W^V.weight，W^O.weight，仅更换输入嵌入矩阵X
- 输入嵌入X做input_norm
- Attention重算
- Attention重算结果不做残差连接直接输出，作为重算后的Attention输出

即：Method1_v3与Method2_v3的差别为：残差连接的修改位点不同，先前层输出重算的截止位置不同（截至MLP输出/截至Attention输出）

Method3

与Method1基本相同，唯一不同之处在于MLP处残差和进行了归一化，且每一层的权重分布为1/m(Method 3.1)或可学习权重(Method3.2)

Method4

与Method2基本相同，唯一不同之处在于Attention处残差和进行了归一化，且每一层的权重分布为1/m(Method 4.1)或可学习权重(Method4.2)

Baseline

```
***** eval metrics *****
epoch = 5.0
eval_accuracy = 0.4966
eval_loss = 2.5789
eval_perplexity = 13.1821
eval_runtime = 0:00:03.86
eval_samples = 143
eval_samples_per_second = 36.969
eval_steps_per_second = 4.653
```

Method1

```
***** eval metrics *****
epoch = 5.0
eval_accuracy = 0.5025
eval_loss = 2.5259
eval_perplexity = 12.5021
eval_runtime = 0:00:06.32
eval_samples = 143
eval_samples_per_second = 22.595
eval_steps_per_second = 2.844
```

Method2

```
***** eval metrics *****
epoch = 5.0
eval_accuracy = 0.4966
eval_loss = 2.5748
eval_perplexity = 13.1291
eval_runtime = 0:00:05.53
eval_samples = 143
eval_samples_per_second = 25.826
eval_steps_per_second = 3.251
```

Method3.1

```
***** eval metrics *****
epoch = 5.0
eval_accuracy = 0.5029
eval_loss = 2.5252
eval_perplexity = 12.494
eval_runtime = 0:00:06.31
eval_samples = 143
eval_samples_per_second = 22.637
eval_steps_per_second = 2.849
```

Method3.2

```
***** eval metrics *****
epoch = 5.0
eval_accuracy = 0.5032
eval_loss = 2.5262
eval_perplexity = 12.5057
eval_runtime = 0:00:06.32
eval_samples = 143
eval_samples_per_second = 22.61
eval_steps_per_second = 2.846
```

Method4.1

```
***** eval metrics *****
epoch = 5.0
eval_accuracy = 0.4979
eval_loss = 2.5658
eval_perplexity = 13.0107
eval_runtime = 0:00:05.56
eval_samples = 143
eval_samples_per_second = 25.714
eval_steps_per_second = 3.237
```

Method4.2

```
***** eval metrics *****
epoch = 5.0
eval_accuracy = 0.4985
eval_loss = 2.5606
eval_perplexity = 12.9437
eval_runtime = 0:00:05.58
eval_samples = 143
eval_samples_per_second = 25.621
eval_steps_per_second = 3.225
```