- 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0：原始Transformer

Method1_1

```
***** eval metrics *****
  epoch                   =        5.0
  eval_accuracy           =     0.4952
  eval_loss               =     2.5836
  eval_perplexity         =    13.2451
  eval_runtime            = 0:00:02.43
  eval_samples            =        143
  eval_samples_per_second =     58.626
  eval_steps_per_second   =      7.379
```

Method1_2

```
***** eval metrics *****
  epoch                   =        5.0
  eval_accuracy           =     0.4955
  eval_loss               =     2.5834
  eval_perplexity         =    13.2421
  eval_runtime            = 0:00:02.42
  eval_samples            =        143
  eval_samples_per_second =     58.859
  eval_steps_per_second   =      7.409
```

Method2_1

```
***** eval metrics *****
  epoch                   =        5.0
  eval_accuracy           =     0.4887
  eval_loss               =     2.6469
  eval_perplexity         =    14.1108
  eval_runtime            = 0:00:02.35
  eval_samples            =        143
  eval_samples_per_second =     60.738
  eval_steps_per_second   =      7.645
```

Method2_2

```
***** eval metrics *****
  epoch                   =        5.0
  eval_accuracy           =     0.4889
  eval_loss               =     2.6457
  eval_perplexity         =     14.094
  eval_runtime            = 0:00:02.37
  eval_samples            =        143
  eval_samples_per_second =     60.262
  eval_steps_per_second   =      7.585
```

Method0_1

```
***** eval metrics *****
  epoch                   =        5.0
  eval_accuracy           =      0.491
  eval_loss               =     2.6246
  eval_perplexity         =    13.7986
  eval_runtime            = 0:00:02.37
  eval_samples            =        143
  eval_samples_per_second =     60.138
  eval_steps_per_second   =       7.57
```

Method0_2

```
***** eval metrics *****
  epoch                   =        5.0
  eval_accuracy           =      0.491
  eval_loss               =     2.6223
  eval_perplexity         =    13.7667
  eval_runtime            = 0:00:02.33
  eval_samples            =        143
  eval_samples_per_second =     61.276
  eval_steps_per_second   =      7.713
```

Conclusion:

性能排序：方法1（改MLP）>方法0（原始Transformer）>方法2（改Attention）