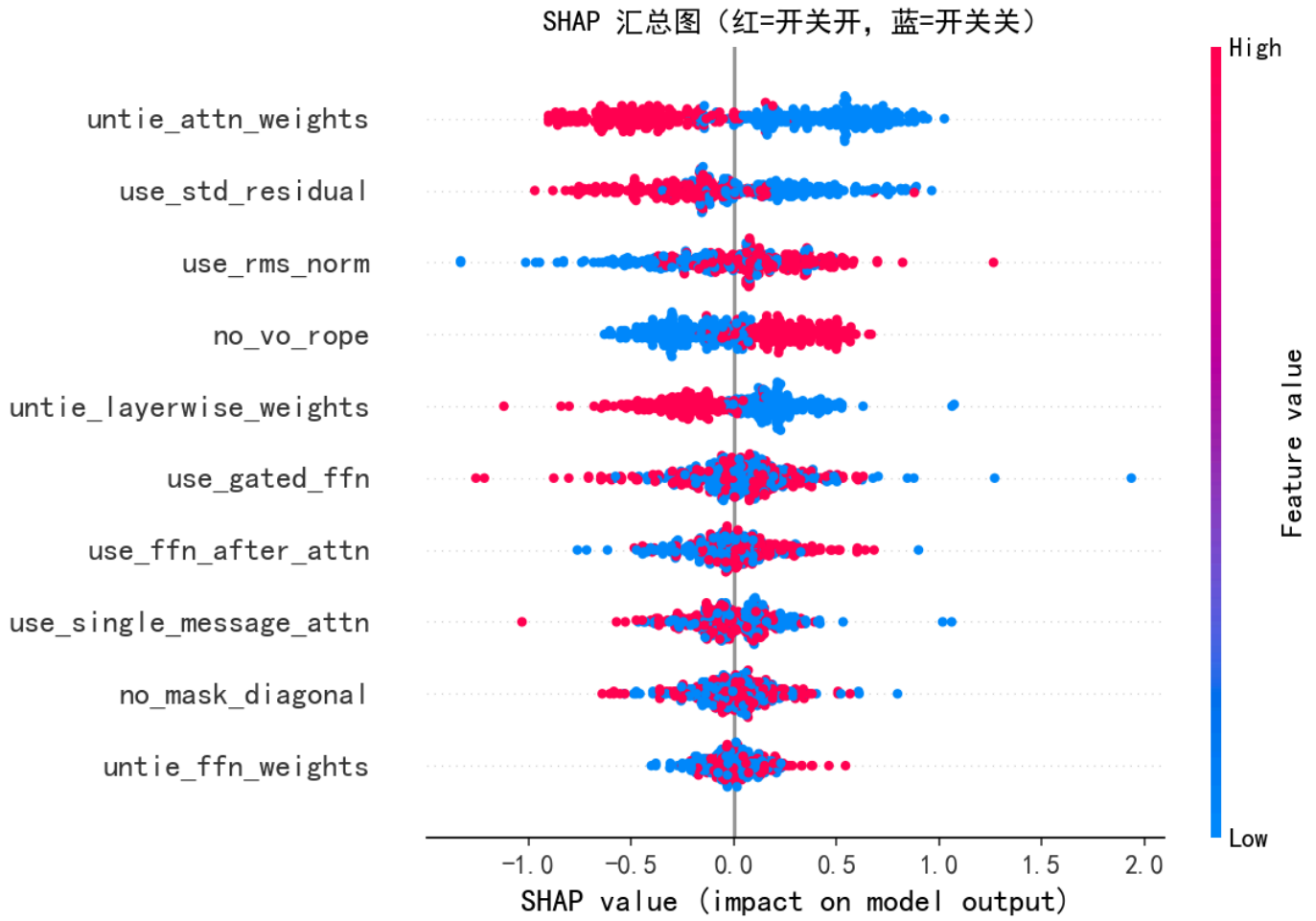
一元分析

使用XGBoost+SHAP方法，分析单个开关开闭情况对于eval/loss的影响值大小。

具体影响值分析

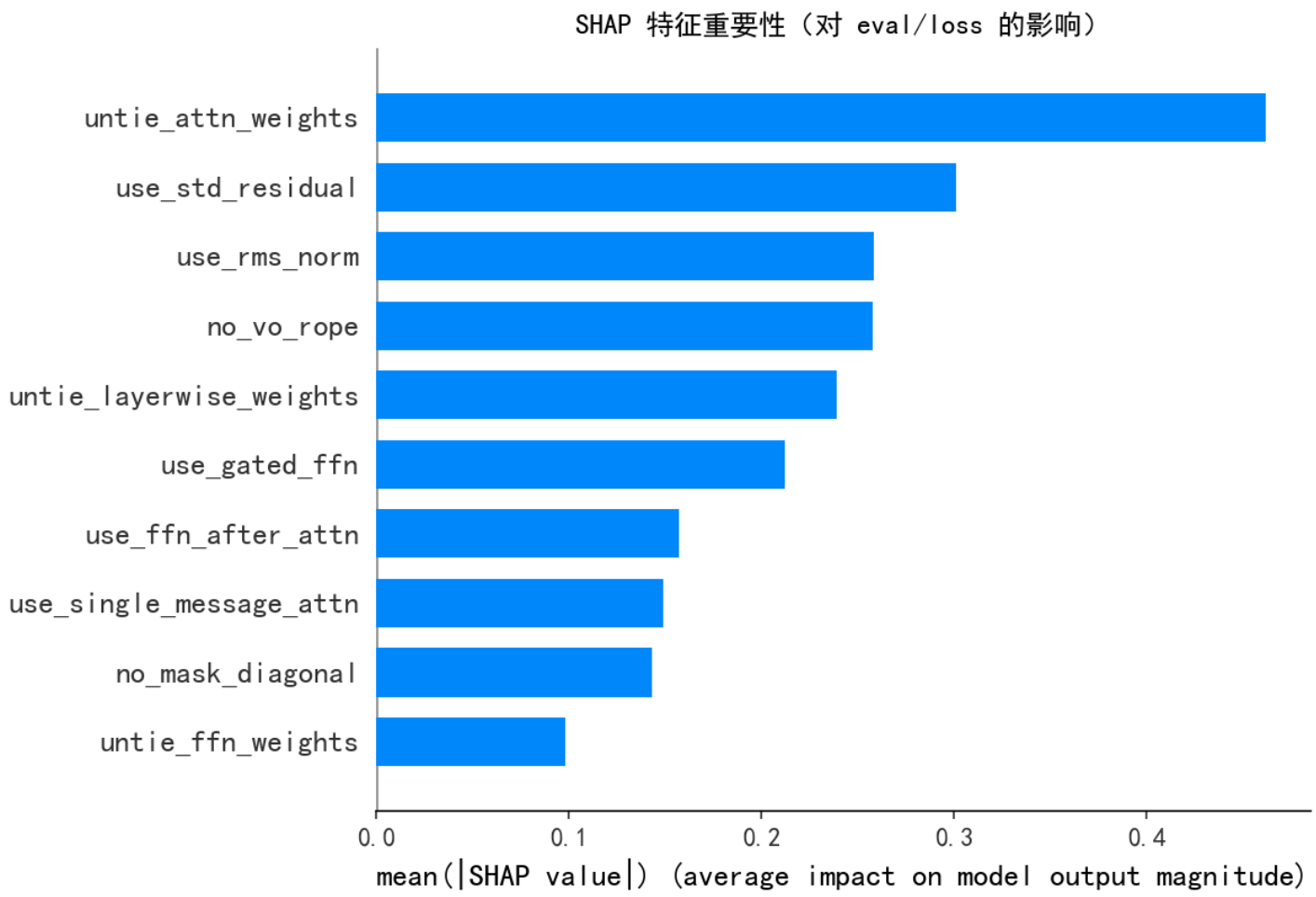


- 一个点为一个实验
- 红色点表示开关打开
- 蓝色点表示开关关闭
- SHAP的原理为拟合出每个实验数据点中单个开关开闭造成的loss改变情况。横轴即为各实验数据点中某个开关开闭对loss的影响

规律分析：

- 前五个开关表现出明显的影响，即：单个开关开闭就会造成显著的loss影响。
- 后五个开关在单开关研究中发现开闭都可能造成loss的正反向不同影响。推断可能是与其他开关存在共同作用，也可能本身影响就不大。

单开关影响强度分析



横轴代表SHAP值的绝对值的平均数，表征单个开关的影响强度。

效果较好实验中单开关开闭情况

最好的5组实验中各开关打开情况占比

	fraction_on
use_std_residual	1
untie_attn_weights	1
use_rms_norm	1
use_gated_ffn	1
untie_ffn_weights	0.8
no_vo_rope	0.8
untie_layerwise_weights	0.8
no_mask_diagonal	0.6
use_ffn_after_attn	0.2
use_single_message_attn	0

最好的20组实验中各开关打开情况占比

	fraction_on
use_rms_norm	1
untie_attn_weights	1
use_std_residual	1
untie_layerwise_weights	0.85
use_gated_ffn	0.8
untie_ffn_weights	0.6
use_single_message_attn	0.55
no_vo_rope	0.55
no_mask_diagonal	0.45
use_ffn_after_attn	0.45

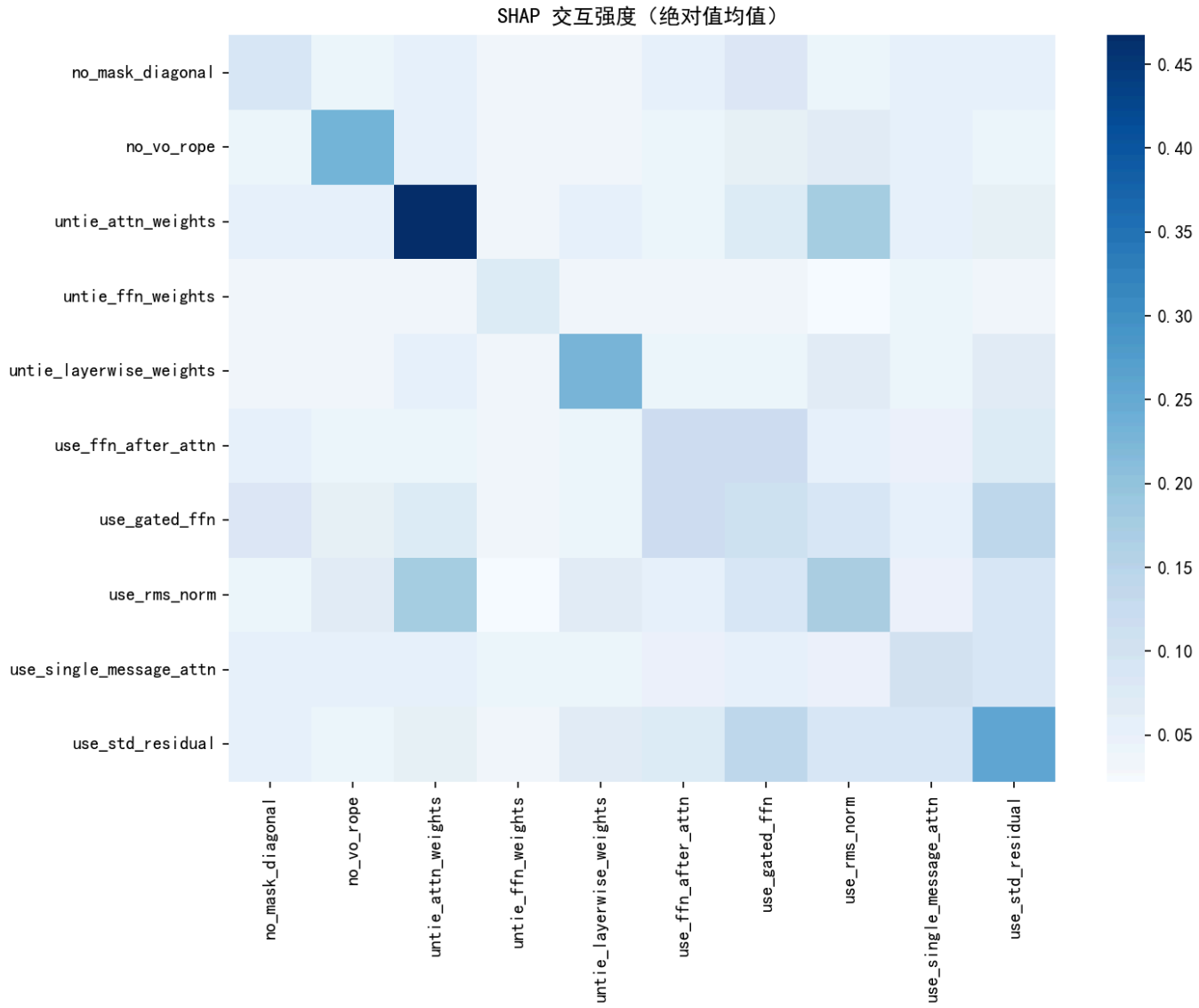
可看出上面柱状图中前三组开关全部打开，其他开关单个开闭情况没有非常明显的规律（存在明显的噪音）。

总结:

- 关注untie_attn_weights & use_std_residual & use_rms_norm（1 2 3号开关。**后续所有开关的编号均以"单开关影响强度分析"柱状图中的次序为准**）开关开闭的影响，这三者存在非常明显的决定作用。
- 部分开关（如后五个开关）红蓝点交错现象非常明显，可能同其他开关存在多元协同作用。需继续分析。

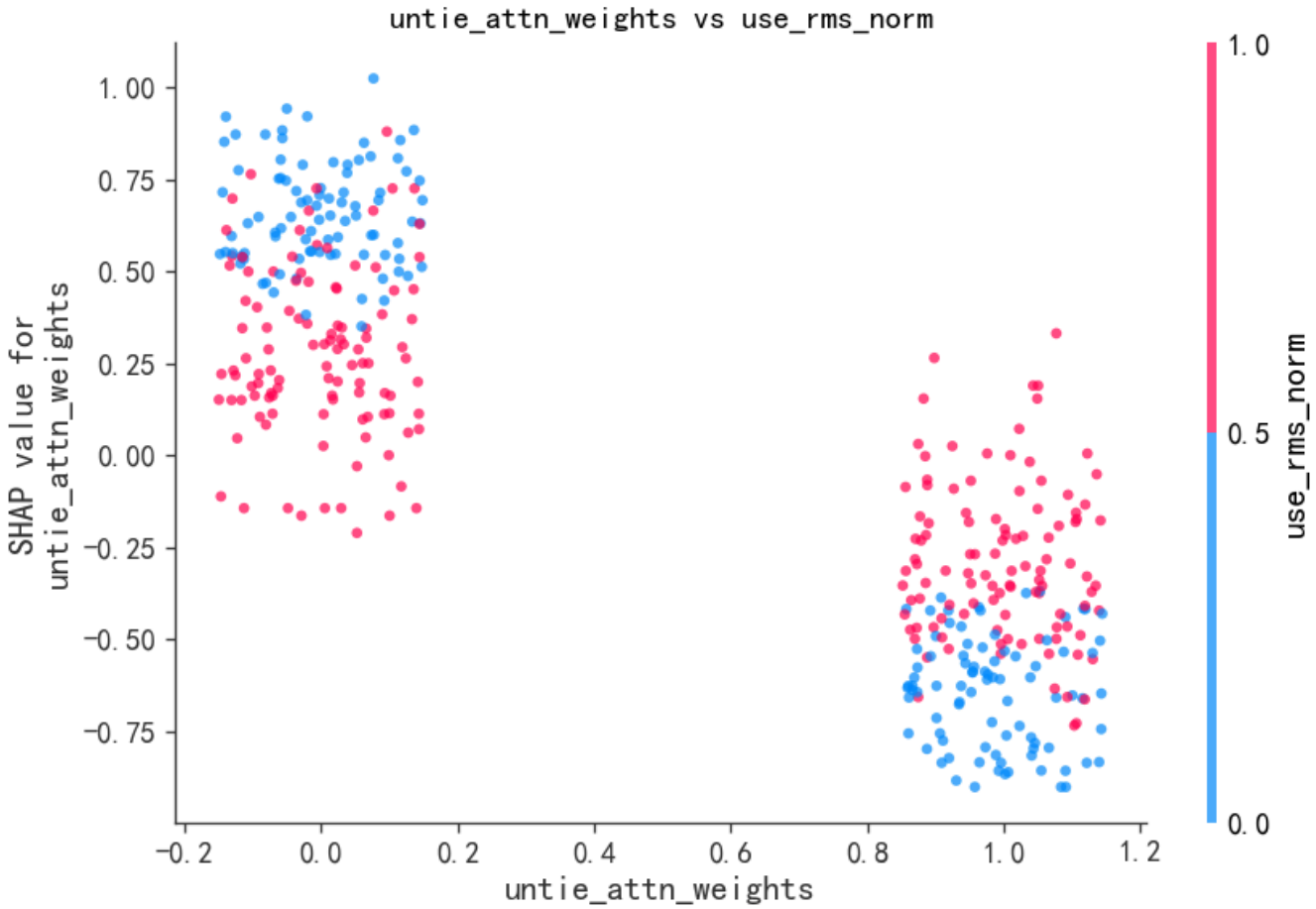
二元分析

二元交互强度分析

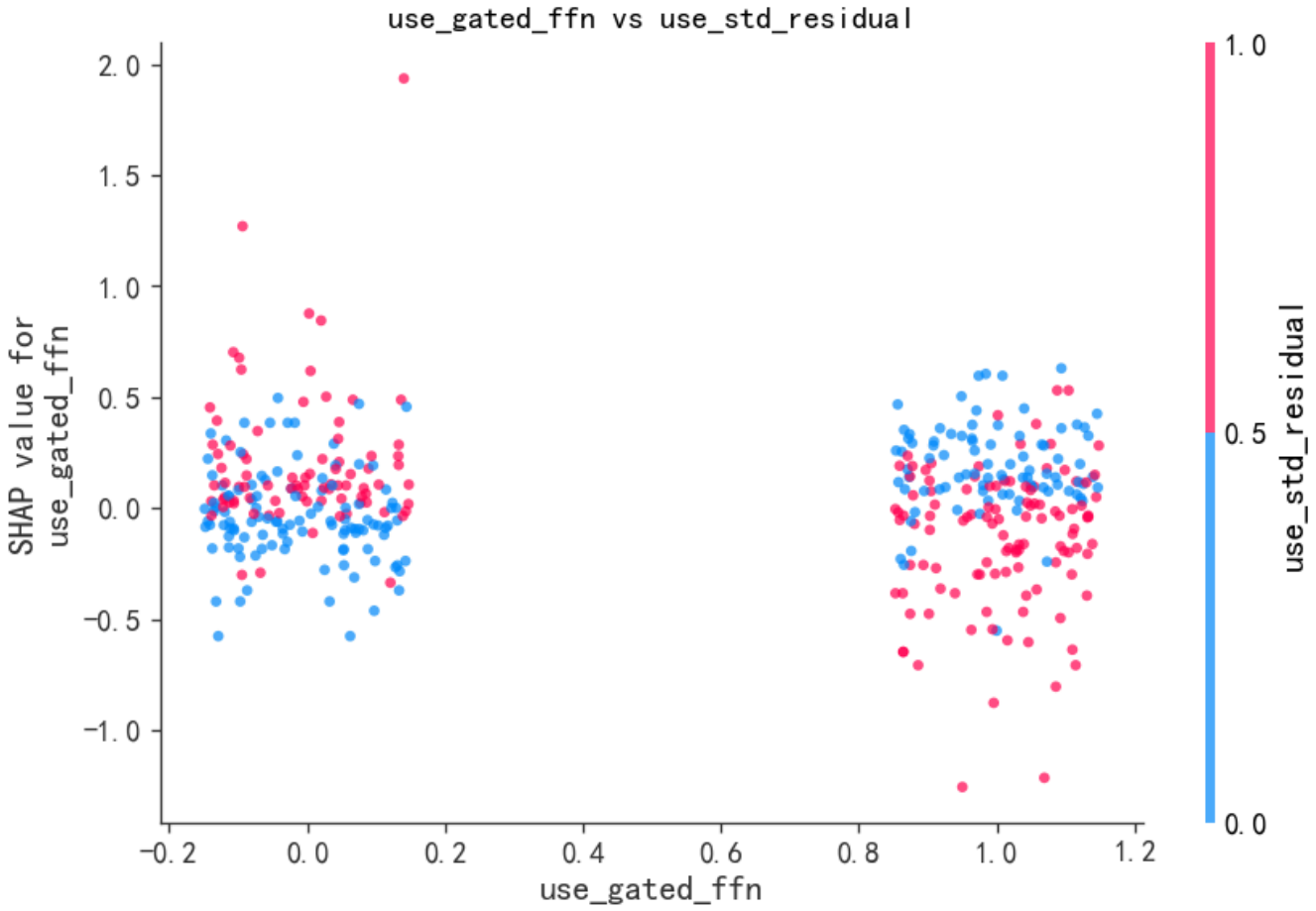


描述了两两之间的交互强度。可看出1 3号开关（即untie_attn_weights & use_rms_norm）、2 6号开关（即use_std_residual & use_gated_ffn）之间存在明显的交互关系。其他两两之间交互关系不大。

具体交互关系如下：**（所有分布图见interaction_analysis文件夹）**



- 横轴为开关1的两个取值（0&1），红蓝色为开关3的两个取值（开或关）。纵轴数值为两个开关取值对loss的影响。
- 从横轴01取值来看，开关1打开对loss下降具有显著影响
- 从红蓝色分布来看：
  - 开关1关闭的情况下，开关3打开（蓝色）会对loss造成一定的下降作用
  - 开关1打开的情况下，开关3关闭（蓝色）会对loss造成一定的下降作用
  - 即开关1 3一闭一开效果略好一些
  - 最差情况为1 3全部关闭，最好情况为1开3闭。二者效果差别较为显著



- 横轴为开关6的两个取值（0&1），红蓝色为开关2的两个取值（开或关）。纵轴数值为两个开关取值对loss的影响。
- 从横轴01取值来看，开关6打开对loss下降影响不大
- 从红蓝色分布来看：
  - 开关6关闭的情况下，开关2关闭（蓝色）似乎会对loss造成一定的下降作用
  - 开关6打开的情况下，开关2打开（蓝色）似乎会对loss造成一定的下降作用，且较为明显
  - 即开关2 6同开同闭效果似乎略好一些
  - 最差情况为6闭2开，最好情况为6开2闭。二者效果略有差别。

开关具体取值情况分析

分析方法：

给定一对开关，共存在四种取值（2*2）。对于每种取值对应的样本共best_count个（如开关1=true&开关3=false的所有样本共best_count=88个）的loss取平均，得best_loss。

按best_loss升序排序，并附带两开关之间的交互强度大小。

（下列数据对应的xlsx表格为interaction_summary_pairs.xlsx）

switch1	switch2	best_loss	best_count	switch1_val	switch2_val	interaction_strength
untie_attn_weights	use_rms_norm	3.503341675	88	TRUE	FALSE	0.179505572
untie_attn_weights	use_std_residual	3.646427598	99	TRUE	TRUE	0.060804833
untie_attn_weights	untie_layerwise_weights	3.683926635	91	TRUE	TRUE	0.050352488
untie_attn_weights	use_ffn_after_attn	3.785446083	94	TRUE	FALSE	0.041582029
untie_attn_weights	use_gated_ffn	3.899271121	104	TRUE	TRUE	0.076150186
no_vo_rope	untie_attn_weights	3.976173608	99	FALSE	TRUE	0.059555165
no_mask_diagonal	untie_attn_weights	3.981052486	101	FALSE	TRUE	0.056339834
untie_attn_weights	untie_ffn_weights	3.986390138	99	TRUE	FALSE	0.033541862
no_vo_rope	use_rms_norm	3.989818608	96	FALSE	FALSE	0.072028965
untie_attn_weights	use_single_message_attn	4.004044096	100	TRUE	FALSE	0.055207148
untie_layerwise_we...	use_std_residual	4.03271874	93	TRUE	TRUE	0.067574069
no_vo_rope	untie_layerwise_weights	4.047769185	89	FALSE	TRUE	0.032936778
use_single_messa...	use_std_residual	4.058193067	97	TRUE	TRUE	0.089132279
use_gated_ffn	use_std_residual	4.063130448	113	TRUE	TRUE	0.143998742
no_vo_rope	use_std_residual	4.070738989	102	FALSE	TRUE	0.040185455
use_ffn_after_attn	use_gated_ffn	4.169859256	90	FALSE	TRUE	0.115938805
use_rms_norm	use_single_message_attn	4.171731574	82	FALSE	FALSE	0.048075657
no_mask_diagonal	use_std_residual	4.172660531	95	FALSE	TRUE	0.055997368
use_rms_norm	use_std_residual	4.192169894	107	TRUE	TRUE	0.094732419
use_ffn_after_attn	use_rms_norm	4.207180285	90	FALSE	FALSE	0.053851824
use_ffn_after_attn	use_std_residual	4.215639023	97	FALSE	TRUE	0.073777504
untie_ffn_weights	use_std_residual	4.219129584	97	TRUE	TRUE	0.035294931
no_vo_rope	use_gated_ffn	4.221182889	101	FALSE	FALSE	0.061407831
no_mask_diagonal	untie_layerwise_weights	4.273015843	93	TRUE	TRUE	0.038921289
no_vo_rope	use_ffn_after_attn	4.277594541	94	FALSE	FALSE	0.042166509
untie_ffn_weights	use_rms_norm	4.29656047	81	TRUE	FALSE	0.022231219
use_gated_ffn	use_rms_norm	4.300692431	86	TRUE	FALSE	0.090307161
untie_layerwise_we...	use_rms_norm	4.306158629	86	FALSE	FALSE	0.063323393
untie_layerwise_we...	use_ffn_after_attn	4.307297795	86	TRUE	FALSE	0.040530533
untie_layerwise_we...	use_single_message_attn	4.310517687	97	TRUE	TRUE	0.041439883
no_mask_diagonal	use_rms_norm	4.320191796	97	FALSE	FALSE	0.044010147
untie_ffn_weights	untie_layerwise_weights	4.32645149	91	FALSE	TRUE	0.033212248
untie_ffn_weights	use_ffn_after_attn	4.327562429	91	FALSE	FALSE	0.038979564
no_vo_rope	untie_ffn_weights	4.33376384	95	FALSE	TRUE	0.030549493
no_mask_diagonal	no_vo_rope	4.334962256	100	FALSE	FALSE	0.040216792
untie_layerwise_we...	use_gated_ffn	4.337715934	99	TRUE	FALSE	0.040210325
use_gated_ffn	use_single_message_attn	4.343388286	94	TRUE	FALSE	0.056962118
no_vo_rope	use_single_message_attn	4.352246128	102	FALSE	TRUE	0.050676491
no_mask_diagonal	use_ffn_after_attn	4.353411838	92	TRUE	FALSE	0.05509191
no_mask_diagonal	use_gated_ffn	4.391486001	113	FALSE	TRUE	0.084126137
use_ffn_after_attn	use_single_message_attn	4.416006695	90	FALSE	TRUE	0.049599618

untie_ffn_weights	use_gated_ffn	4.440440613	98	TRUE	TRUE	0.032801181
no_mask_diagonal	use_single_message_attn	4.459104673	101	FALSE	FALSE	0.059234679
no_mask_diagonal	untie_ffn_weights	4.505983088	101	FALSE	FALSE	0.033720713
untie_ffn_weights	use_single_message_attn	4.518281231	97	FALSE	TRUE	0.042942997

可看出：

- 开关attention_weights（开关1）在loss最低的几个样本中几乎全部打开。
- 开关1开3闭（即交互强度最大的一组）效果最为显著，明显好于其他组别。

总结

- 开关两两关系中交互作用最显著的是开关（1 3）与开关（2 6）组合。其中开关1 3组合的交互作用最强。除去这两组外，很难找到有明显交互作用的组别。
- 可能需要重点关注一下开关1 3之间的关系。一元分析时发现1和3作为单个开关就会对loss的变换造成显著影响，二元分析时发现1 3之间还存在较强的相互作用。

三元分析（作为参考）

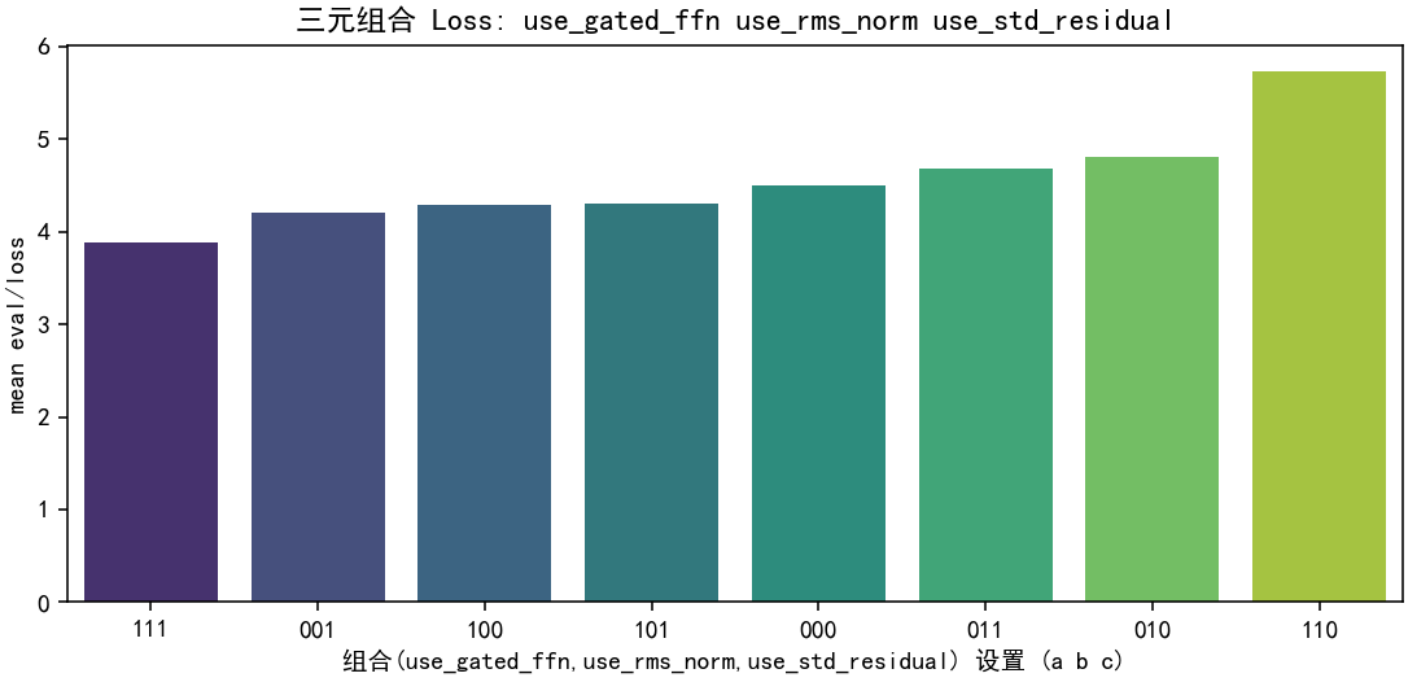
三元交互强度分析

下图为交互强度靠前的部分组别。完整数据见"three_way_model_interactions.xlsx"文件。

switch1	switch2	switch3	model_interaction_strength
use_gated_ffn	use_rms_norm	use_std_residual	0.198519533
use_ffn_after_attn	use_gated_ffn	use_rms_norm	0.198216474
untie_attn_weights	use_rms_norm	use_std_residual	0.18307325
no_mask_diagonal	no_vo_rope	use_single_message_attn	0.182410218
no_vo_rope	untie_attn_weights	use_rms_norm	0.182224163
untie_attn_weights	use_gated_ffn	use_rms_norm	0.158314337
untie_attn_weights	untie_layerwise_weights	use_rms_norm	0.156304492
no_vo_rope	untie_layerwise_weights	use_single_message_attn	0.153460737
no_mask_diagonal	use_gated_ffn	use_std_residual	0.151618584
no_mask_diagonal	use_gated_ffn	use_rms_norm	0.1506587
untie_layerwise_we...	use_rms_norm	use_std_residual	0.150611137
untie_ffn_weights	use_single_message_attn	use_std_residual	0.149392569
untie_attn_weights	use_single_message_attn	use_std_residual	0.139360273
untie_ffn_weights	use_ffn_after_attn	use_std_residual	0.131969574
no_vo_rope	use_ffn_after_attn	use_gated_ffn	0.131677921
use_ffn_after_attn	use_single_message_attn	use_std_residual	0.130598108
use_gated_ffn	use_single_message_attn	use_std_residual	0.129454116
no_mask_diagonal	no_vo_rope	untie_layerwise_weights	0.12869921
untie_ffn_weights	use_ffn_after_attn	use_single_message_attn	0.128465625
untie_attn_weights	use_gated_ffn	use_std_residual	0.124815264
no_vo_rope	use_ffn_after_attn	use_single_message_attn	0.124494881
use_gated_ffn	use_rms_norm	use_single_message_attn	0.124203716
no_mask_diagonal	untie_layerwise_weights	use_single_message_attn	0.124116186
untie_layerwise_we...	use_ffn_after_attn	use_std_residual	0.123421668
no_mask_diagonal	use_ffn_after_attn	use_std_residual	0.121942073
untie_attn_weights	use_rms_norm	use_single_message_attn	0.121895862
untie_ffn_weights	use_gated_ffn	use_single_message_attn	0.116714925
no_vo_rope	use_rms_norm	use_std_residual	0.11627924
no_vo_rope	untie_attn_weights	use_single_message_attn	0.114456533
untie_attn_weights	use_ffn_after_attn	use_single_message_attn	0.112569538
no_vo_rope	use_single_message_attn	use_std_residual	0.111490971
no_vo_rope	untie_attn_weights	use_std_residual	0.110961721

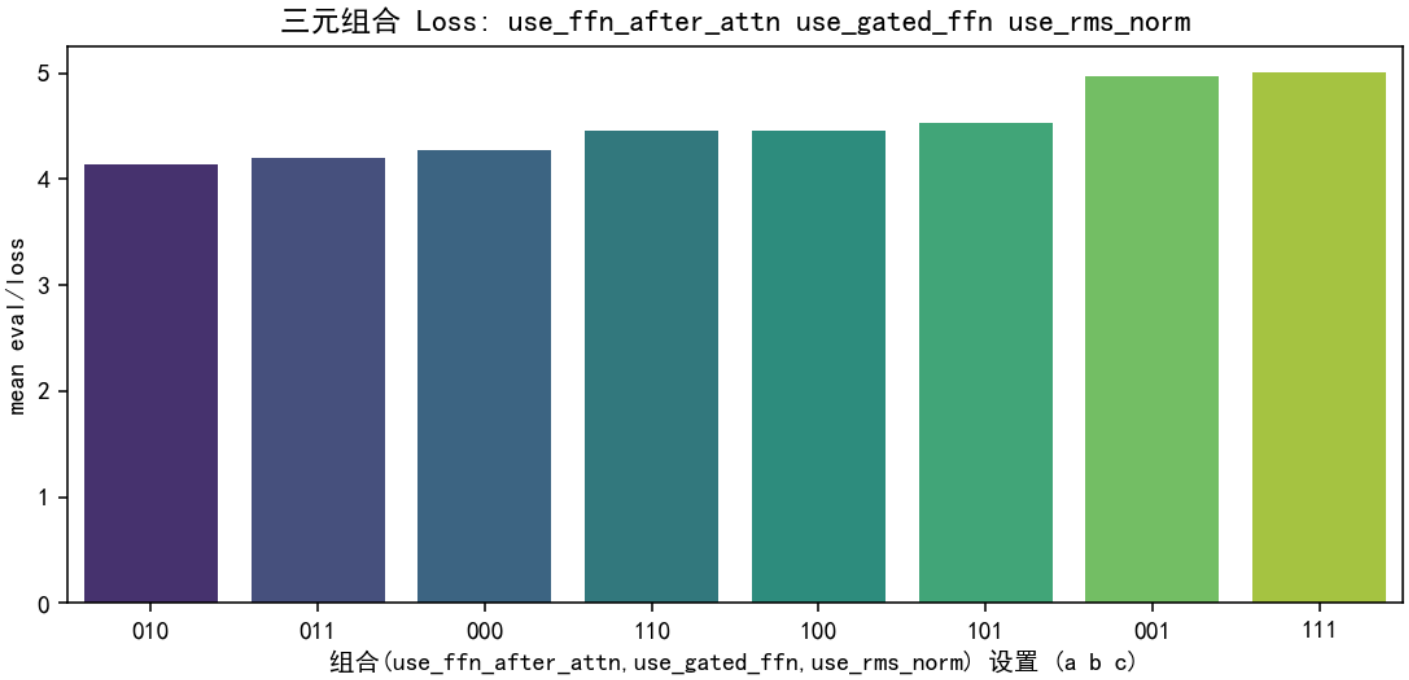
use_rms_norm	use_single_message_attn	use_std_residual	0.110944269
untie_attn_weights	use_ffn_after_attn	use_rms_norm	0.110318153
untie_attn_weights	untie_layerwise_weights	use_gated_ffn	0.109544976
no_vo_rope	untie_ffn_weights	use_single_message_attn	0.109073998
no_vo_rope	use_gated_ffn	use_rms_norm	0.108934377
no_vo_rope	use_ffn_after_attn	use_std_residual	0.108915481

第一组具体分布：**（所有分布图见interaction_analysis文件夹）**



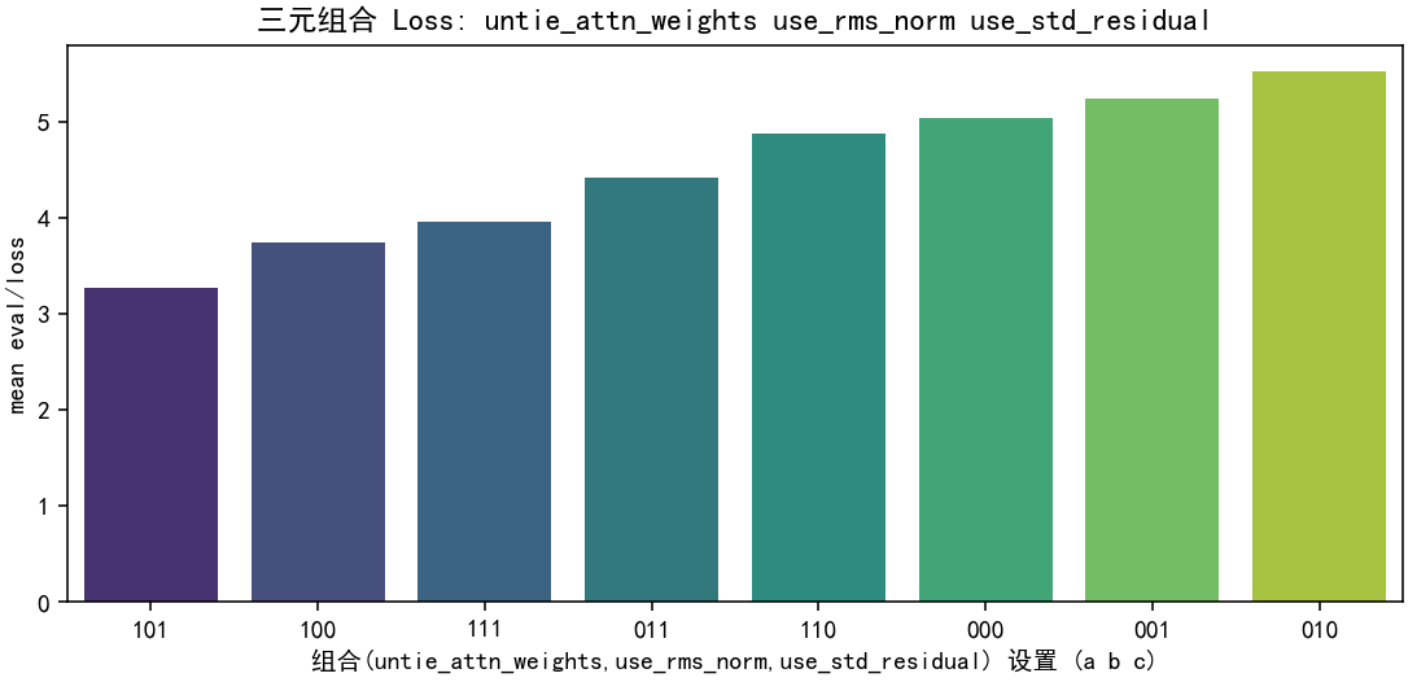
- 这一组的交互系数最大，但看起来刨去最低值和最高值之外，相差并不显著。而最低值和最高值之间仅仅差了use_std_residual一个开关，但导致了显著的极差。
- 也许可以推断出：use_std_residual开关可能发挥重要的作用，但前提是use_gated_ffn与use_rms_norm全部打开。

第二组具体分布：



- 这一组好像相差不算很大。如何分析?

第三组具体分布:



- 这一组正好对应了1 2 3号开关，即单开关影响最大的三个。可以看出八个数值之间都有显著的差别。
- 另外一个特殊的现象是，三个开关的01取值完全对称，也就是101与010正好是最好最差，100与001正好是次好次差等等。
- 如何解释这一现象?

开关具体取值情况（参考）

完整结果见文件"interaction_summary_three.xlsx"

switch1	switch2	switch3	best_loss	best_count	switch1_val	switch2_val	switch3_val	interaction_strength
untie_attn_weights	untie_layerwise_weights	use_std_residual	3.235960...	44	TRUE	TRUE	TRUE	0.059577126
untie_attn_weights	use_rms_norm	use_std_residual	3.270496...	45	TRUE	FALSE	TRUE	0.111680947
untie_attn_weights	use_rms_norm	use_single_message_attn	3.359429...	44	TRUE	FALSE	FALSE	0.094262786
untie_attn_weights	use_gated_ffn	use_rms_norm	3.40112747	40	TRUE	FALSE	FALSE	0.115320981
untie_attn_weights	use_gated_ffn	use_std_residual	3.402527...	57	TRUE	TRUE	TRUE	0.093651257
untie_attn_weights	use_ffn_after_attn	use_gated_ffn	3.423967...	49	TRUE	FALSE	TRUE	0.077890337
untie_attn_weights	use_single_message_attn	use_std_residual	3.427768...	52	TRUE	TRUE	TRUE	0.068381421
untie_attn_weights	use_ffn_after_attn	use_rms_norm	3.434469...	46	TRUE	FALSE	FALSE	0.09164647
untie_attn_weights	untie_ffn_weights	use_rms_norm	3.470475...	40	TRUE	TRUE	FALSE	0.07842622
untie_attn_weights	untie_layerwise_weights	use_rms_norm	3.476094...	39	TRUE	TRUE	FALSE	0.09772715
no_vo_rope	untie_attn_weights	use_rms_norm	3.493612...	50	FALSE	TRUE	FALSE	0.103696562
no_mask_diagonal	untie_attn_weights	use_rms_norm	3.495318...	47	FALSE	TRUE	FALSE	0.093285181
no_vo_rope	untie_attn_weights	use_std_residual	3.519854...	49	FALSE	TRUE	TRUE	0.053515151
untie_attn_weights	untie_layerwise_weights	use_ffn_after_attn	3.578729...	41	TRUE	TRUE	FALSE	0.044155017
no_mask_diagonal	untie_attn_weights	use_std_residual	3.587057...	50	TRUE	TRUE	TRUE	0.057714012
no_vo_rope	untie_attn_weights	untie_layerwise_weights	3.588068...	43	FALSE	TRUE	TRUE	0.047614809
untie_attn_weights	untie_ffn_weights	use_std_residual	3.595035...	49	TRUE	TRUE	TRUE	0.043213874
untie_attn_weights	untie_ffn_weights	untie_layerwise_weights	3.623228...	45	TRUE	TRUE	TRUE	0.039035533
untie_attn_weights	use_ffn_after_attn	use_std_residual	3.643488...	50	TRUE	FALSE	TRUE	0.058721453
untie_attn_weights	untie_layerwise_weights	use_single_message_attn	3.649717...	48	TRUE	TRUE	FALSE	0.048999835
no_mask_diagonal	untie_attn_weights	untie_layerwise_weights	3.650052...	47	TRUE	TRUE	TRUE	0.048537869
untie_attn_weights	untie_layerwise_weights	use_gated_ffn	3.654146...	46	TRUE	TRUE	FALSE	0.055571001
no_mask_diagonal	untie_attn_weights	use_ffn_after_attn	3.662861...	45	TRUE	TRUE	FALSE	0.051004592
untie_attn_weights	untie_ffn_weights	use_gated_ffn	3.685716...	50	TRUE	TRUE	TRUE	0.047497746
no_vo_rope	untie_attn_weights	use_gated_ffn	3.690139...	51	TRUE	TRUE	TRUE	0.065704398
untie_layerwise_we...	use_rms_norm	use_std_residual	3.701679...	46	TRUE	TRUE	TRUE	0.07520996
untie_attn_weights	use_ffn_after_attn	use_single_message_attn	3.716799...	42	TRUE	FALSE	TRUE	0.048796266
no_vo_rope	use_single_message_attn	use_std_residual	3.753908...	44	FALSE	TRUE	TRUE	0.059998076
no_vo_rope	untie_attn_weights	use_ffn_after_attn	3.763071...	42	FALSE	TRUE	FALSE	0.047767904
use_rms_norm	use_single_message_attn	use_std_residual	3.763627...	52	TRUE	TRUE	TRUE	0.077313453
untie_attn_weights	use_gated_ffn	use_single_message_attn	3.767538...	53	TRUE	TRUE	FALSE	0.062773153
untie_attn_weights	untie_ffn_weights	use_ffn_after_attn	3.773730...	51	TRUE	TRUE	FALSE	0.038034484
no_vo_rope	untie_layerwise_weights	use_std_residual	3.798351...	48	FALSE	TRUE	TRUE	0.046898767
no_mask_diagonal	untie_attn_weights	use_single_message_attn	3.802190...	51	FALSE	TRUE	FALSE	0.056927219
use_ffn_after_attn	use_gated_ffn	use_std_residual	3.835804...	53	FALSE	TRUE	TRUE	0.111238353
no_vo_rope	untie_layerwise_weights	use_gated_ffn	3.838536...	45	FALSE	TRUE	FALSE	0.044851646
untie_layerwise_we...	use_single_message_attn	use_std_residual	3.839406...	46	TRUE	TRUE	TRUE	0.066048741
no_vo_rope	use_gated_ffn	use_rms_norm	3.851585...	45	FALSE	FALSE	FALSE	0.074581318

观测结果及疑点总结

观测结果

- 开关1 2 3作为单个开关时就会有显著的影响（无需考虑其他开关情况）
- 开关组1&3、开关组2&6之间相互影响较为显著。尤其是开关组1&3
- 开关组1&2&3在三元分析时也具有显著的影响，也许可以重点研究开关1、2、3相关的性质?

疑点

如何解释三元分析中的第二、第三组的数据，尤其是第三组（开关1&2&3）的相关规律?