**实验（实习）报告**

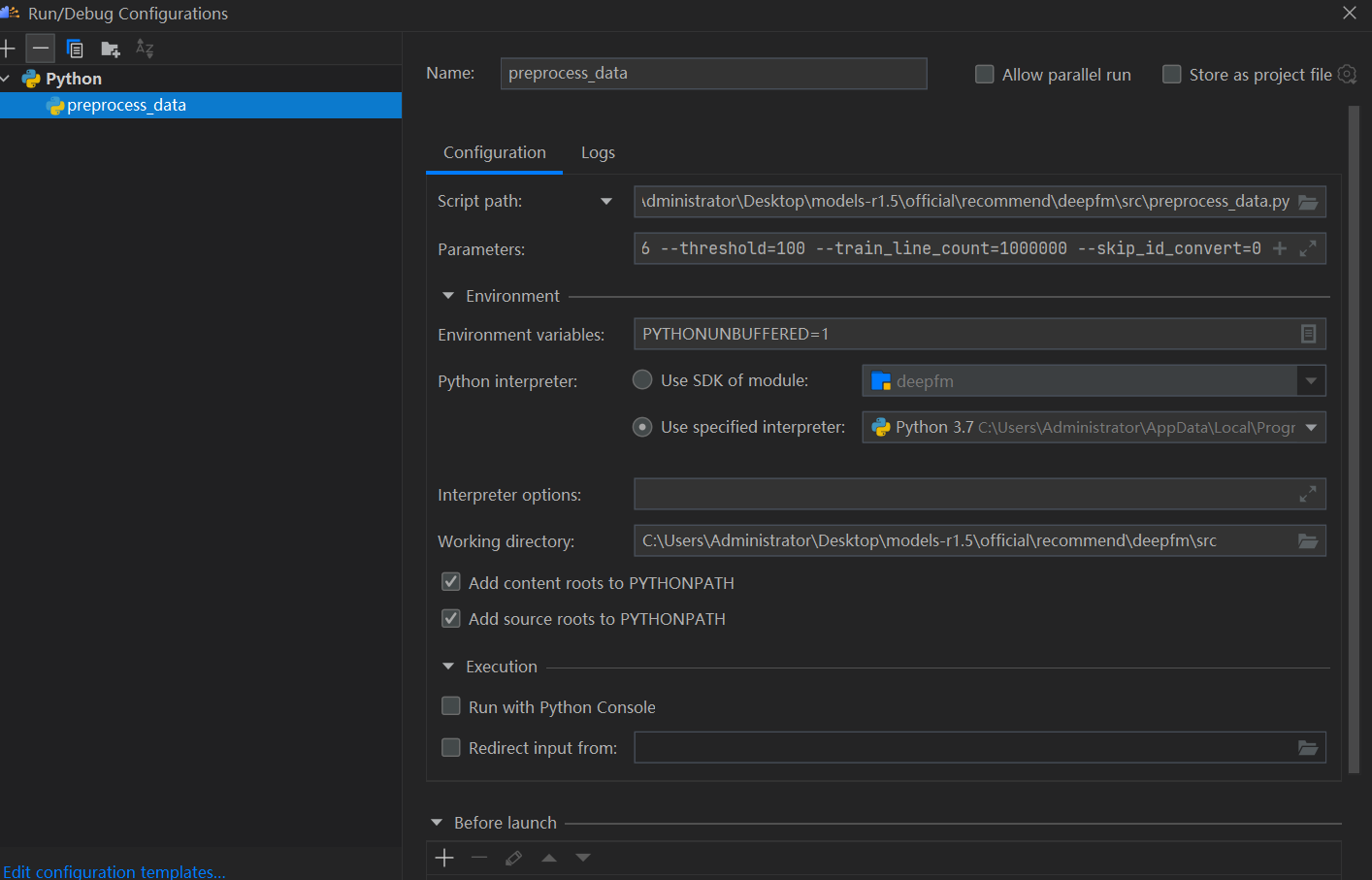
1. 实验目的
2. 掌握Wide&Deep算法的基本原理以及点击率预测的实验流程。
3. 实验任务

基于Criteo数据，使用推荐系统的深度学习算法wide&deep实现广告推荐

1. 实验步骤
2. 数据集预处理

步骤 1使用MindStudio打开deepfm/src/preprocess\_data.py脚本文件

步骤 2设置preprocess\_data.py脚本参数



步骤 3运行preprocess\_data.py脚本文件

{'enable\_modelarts': 'Whether training on modelarts, default: False', 'data\_url': 'Dataset url for obs', 'train\_url': 'Training output url for obs', 'data\_path': 'Dataset path for local', 'output\_path': 'Training output path for local', 'device\_target': 'device target, support Ascend, GPU and CPU.', 'dataset\_path': 'Dataset

{'enable\_modelarts': 'Whether training on modelarts, default: False', 'data\_url': 'Dataset url for obs', 'train\_url': 'Training output url for obs', 'data\_path': 'Dataset path for local', 'output\_path': 'Training output path for local', 'device\_target': 'device target, support Ascend, GPU and CPU.', 'dataset\_path': 'Dataset path', 'batch\_size': 'batch size', 'ckpt\_path': 'Checkpoint path', 'eval\_file\_name': 'Auc log file path. Default: "./auc.log"', 'loss\_file\_name': 'Loss log file path. Default: "./loss.log"', 'do\_eval': 'Do evaluation or not, only support "True" or "False". Default: "True"', 'checkpoint\_path': 'Checkpoint file path', 'device\_id': 'Device id', 'ckpt\_file': 'Checkpoint file path.', 'file\_name': 'output file name.', 'file\_format': 'file format', 'result\_path': 'Result path', 'label\_path': 'label path', 'dense\_dim': 'The number of your continues fields', 'slot\_dim': 'The number of your sparse fields, it can also be called catelogy features.', 'threshold': 'Word frequency below this will be regarded as OOV. It aims to reduce the vocab size', 'train\_line\_count': 'The number of examples in your dataset', 'skip\_id\_convert': 'Skip the id convert, regarding the original id as the final id.'}

{'batch\_size': 16000,

 'checkpoint\_path': '/cache/train/deepfm-5\_2582.ckpt',

 'checkpoint\_url': '',

 'ckpt\_file': '/cache/train/deepfm-5\_2582.ckpt',

 'ckpt\_file\_name\_prefix': 'deepfm',

 'ckpt\_path': '/cache/train',

 'config\_path': 'C:\\Users\\Administrator\\Desktop\\models-r1.5\\official\\recommend\\deepfm\\src\\model\_utils\\../../default\_config.yaml',

 'convert\_dtype': True,

 'data\_emb\_dim': 80,

 'data\_field\_size': 39,

 'data\_format': 1,

 'data\_path': '../data/',

 'data\_url': '',

 'data\_vocab\_size': 184965,

 'dataset\_path': '/cache/data',

 'deep\_layer\_args': [[1024, 512, 256, 128], 'relu'],

 'dense\_dim': 13,

 'device\_id': 0,

 'device\_target': 'Ascend',

 'do\_eval': 'True',

 'enable\_modelarts': False,

 'enable\_profiling': False,

 'epsilon': 5e-08,

 'eval\_callback': True,

 'eval\_file\_name': './auc.log',

 'file\_format': 'AIR',

 'file\_name': 'deepfm',

 'init\_args': [-0.01, 0.01],

 'keep\_checkpoint\_max': 50,

 'keep\_prob': 0.9,

 'l2\_coef': 8e-05,

 'label\_path': '',

 'learning\_rate': 0.0005,

 'load\_path': '/cache/checkpoint\_path',

 'loss\_callback': True,

 'loss\_file\_name': './loss.log',

 'loss\_scale': 1024.0,

 'output\_path': '/cache/train',

 'result\_path': './preprocess\_Result',

 'save\_checkpoint': True,

 'save\_checkpoint\_steps': 1,

 'skip\_id\_convert': 0,

 'slot\_dim': 26,

 'test\_num\_of\_parts': 3,

 'threshold': 100,

 'train\_epochs': 5,

 'train\_line\_count': 1000000,

 'train\_num\_of\_parts': 21,

 'train\_url': '',

 'weight\_bias\_init': ['normal', 'normal']}

Please check the above information for the configurations

Have handled 100w lines.

val\_max\_dict.items()[:50]:[('val\_1', 780.0), ('val\_2', 18522.0), ('val\_3', 65535.0), ('val\_4', 507.0), ('val\_5', 2527030.0), ('val\_6', 233523.0), ('val\_7', 8807.0), ('val\_8', 5064.0), ('val\_9', 19327.0), ('val\_10', 8.0), ('val\_11', 147.0), ('val\_12', 768.0), ('val\_13', 6702.0)]

val\_min\_dict.items()[:50]:[('val\_1', 0), ('val\_2', -2.0), ('val\_3', 0), ('val\_4', 0), ('val\_5', 0), ('val\_6', 0), ('val\_7', 0), ('val\_8', 0), ('val\_9', 0), ('val\_10', 0), ('val\_11', 0), ('val\_12', 0), ('val\_13', 0)]

cat2id\_dict.size:14685

cat2id.dict.items()[:50]:[('val\_1', 0), ('val\_2', 1), ('val\_3', 2), ('val\_4', 3), ('val\_5', 4), ('val\_6', 5), ('val\_7', 6), ('val\_8', 7), ('val\_9', 8), ('val\_10', 9), ('val\_11', 10), ('val\_12', 11), ('val\_13', 12), ('OOVcat\_1', 13), ('OOVcat\_2', 14), ('OOVcat\_3', 15), ('OOVcat\_4', 16), ('OOVcat\_5', 17), ('OOVcat\_6', 18), ('OOVcat\_7', 19), ('OOVcat\_8', 20), ('OOVcat\_9', 21), ('OOVcat\_10', 22), ('OOVcat\_11', 23), ('OOVcat\_12', 24), ('OOVcat\_13', 25), ('OOVcat\_14', 26), ('OOVcat\_15', 27), ('OOVcat\_16', 28), ('OOVcat\_17', 29), ('OOVcat\_18', 30), ('OOVcat\_19', 31), ('OOVcat\_20', 32), ('OOVcat\_21', 33), ('OOVcat\_22', 34), ('OOVcat\_23', 35), ('OOVcat\_24', 36), ('OOVcat\_25', 37), ('OOVcat\_26', 38), ('cat\_1\_68fd1e64', 39), ('cat\_1\_287e684f', 40), ('cat\_1\_8cf07265', 41), ('cat\_1\_05db9164', 42), ('cat\_1\_439a44a4', 43), ('cat\_1\_241546e0', 44), ('cat\_1\_be589b51', 45), ('cat\_1\_5a9ed9b0', 46), ('cat\_1\_3c9d8785', 47), ('cat\_1\_1464facd', 48), ('cat\_1\_7e5c2ff4', 49)]

all\_indices.size:1000000

test\_indices\_set.size:100000

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Have handle 100w lines.

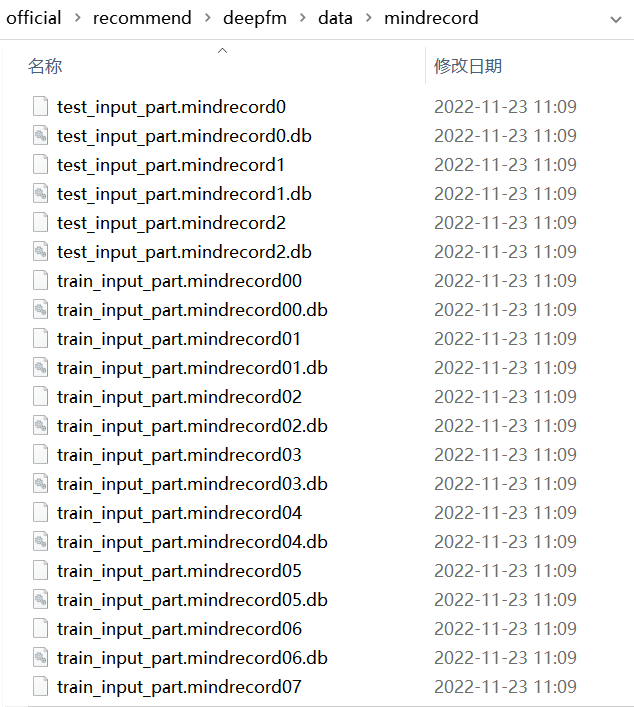
----------------------------------------------------------------------------------------------------------------------------------

items\_error\_size\_lineCount.size(): 0.

----------------------------------------------------------------------------------------------------------------------------------

Process finished with exit code 0

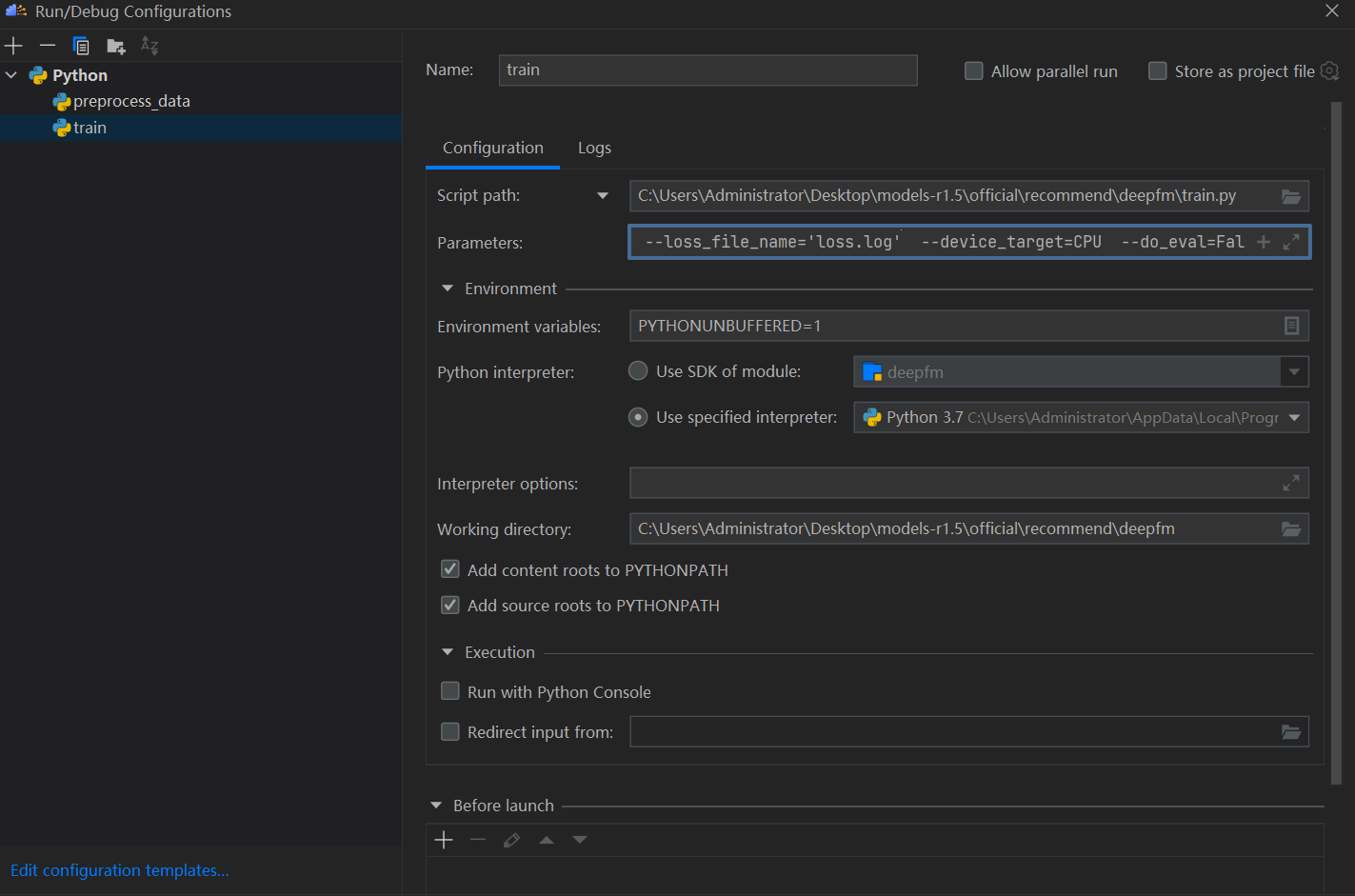
步骤 4查看MindRecord文件



1. 训练过程

步骤 1打开deepfm\train.py脚本文件

步骤 2设置train.py脚本参数



步骤 3运行train.py脚本文件

{'enable\_modelarts': 'Whether training on modelarts, default: False', 'data\_url': 'Dataset url for obs', 'train\_url': 'Training output url for obs', 'data\_path': 'Dataset path for local', 'output\_path': 'Training output path for local', 'device\_target': 'device target, support Ascend, GPU and CPU.', 'dataset\_path': 'Dataset path', 'batch\_size': 'batch size', 'ckpt\_path': 'Checkpoint path', 'eval\_file\_name': 'Auc log file path. Default: "./auc.log"', 'loss\_file\_name': 'Loss log file path. Default: "./loss.log"', 'do\_eval': 'Do evaluation or not, only support "True" or "False". Default: "True"', 'checkpoint\_path': 'Checkpoint file path', 'device\_id': 'Device id', 'ckpt\_file': 'Checkpoint file path.', 'file\_name': 'output file name.', 'file\_format': 'file format', 'result\_path': 'Result path', 'label\_path': 'label path', 'dense\_dim': 'The number of your continues fields', 'slot\_dim': 'The number of your sparse fields, it can also be called catelogy features.', 'threshold': 'Word frequency below this will be regarded as OOV. It aims to reduce the vocab size', 'train\_line\_count': 'The number of examples in your dataset', 'skip\_id\_convert': 'Skip the id convert, regarding the original id as the final id.'}

{'batch\_size': 16000,

 'checkpoint\_path': '/cache/train/deepfm-5\_2582.ckpt',

 'checkpoint\_url': '',

 'ckpt\_file': '/cache/train/deepfm-5\_2582.ckpt',

 'ckpt\_file\_name\_prefix': 'deepfm',

 'ckpt\_path': 'checkpoint',

 'config\_path': 'C:\\Users\\Administrator\\Desktop\\models-r1.5\\official\\recommend\\deepfm\\src\\model\_utils\\../../default\_config.yaml',

 'convert\_dtype': True,

 'data\_emb\_dim': 80,

 'data\_field\_size': 39,

 'data\_format': 1,

 'data\_path': '/cache/data',

 'data\_url': '',

 'data\_vocab\_size': 184965,

 'dataset\_path': './data/mindrecord/',

 'deep\_layer\_args': [[1024, 512, 256, 128], 'relu'],

 'dense\_dim': 13,

 'device\_id': 0,

 'device\_target': 'CPU',

 'do\_eval': 'False',

 'enable\_modelarts': False,

 'enable\_profiling': False,

 'epsilon': 5e-08,

 'eval\_callback': True,

 'eval\_file\_name': "'auc.log'",

 'file\_format': 'AIR',

 'file\_name': 'deepfm',

 'init\_args': [-0.01, 0.01],

 'keep\_checkpoint\_max': 50,

 'keep\_prob': 0.9,

 'l2\_coef': 8e-05,

 'label\_path': '',

 'learning\_rate': 0.0005,

 'load\_path': '/cache/checkpoint\_path',

 'loss\_callback': True,

 'loss\_file\_name': "'loss.log'",

 'loss\_scale': 1024.0,

 'output\_path': '/cache/train',

 'result\_path': './preprocess\_Result',

 'save\_checkpoint': True,

 'save\_checkpoint\_steps': 1,

 'skip\_id\_convert': 0,

 'slot\_dim': 26,

 'test\_num\_of\_parts': 3,

 'threshold': 100,

 'train\_epochs': 5,

 'train\_line\_count': 45840617,

 'train\_num\_of\_parts': 21,

 'train\_url': '',

 'weight\_bias\_init': ['normal', 'normal']}

Please check the above information for the configurations

[WARNING] ME(15172:15340,MainProcess):2022-11-23-12:02:56.301.107 [mindspore\train\model.py:500] The CPU cannot support dataset sink mode currently.So the training process will be performed with dataset not sink.

epoch: 1 step: 1, loss is 0.7508878111839294

epoch: 1 step: 2, loss is 0.7417744994163513

epoch: 1 step: 3, loss is 0.7332229018211365

epoch: 1 step: 4, loss is 0.7265874743461609

epoch: 1 step: 5, loss is 0.7198621034622192

epoch: 1 step: 6, loss is 0.7143847346305847

epoch: 1 step: 7, loss is 0.7062808275222778

epoch: 1 step: 8, loss is 0.7011118531227112

epoch: 1 step: 9, loss is 0.6957386136054993

epoch: 1 step: 10, loss is 0.6884331107139587

...

epoch: 5 step: 53, loss is 0.46436214447021484

epoch: 5 step: 54, loss is 0.47739702463150024

epoch: 5 step: 55, loss is 0.4688734710216522

epoch time: 481341.560 ms, per step time: 8751.665 ms

Process finished with exit code 0

步骤 4查看loss.log日志

2022-11-23 12:03:24 epoch: 1 step: 1, loss is 0.7508878111839294

2022-11-23 12:03:33 epoch: 1 step: 2, loss is 0.7417744994163513

2022-11-23 12:03:41 epoch: 1 step: 3, loss is 0.7332229018211365

2022-11-23 12:03:49 epoch: 1 step: 4, loss is 0.7265874743461609

2022-11-23 12:03:57 epoch: 1 step: 5, loss is 0.7198621034622192

2022-11-23 12:04:06 epoch: 1 step: 6, loss is 0.7143847346305847

2022-11-23 12:04:15 epoch: 1 step: 7, loss is 0.7062808275222778

2022-11-23 12:04:22 epoch: 1 step: 8, loss is 0.7011118531227112

2022-11-23 12:04:29 epoch: 1 step: 9, loss is 0.6957386136054993

2022-11-23 12:04:37 epoch: 1 step: 10, loss is 0.6884331107139587

2022-11-23 12:04:45 epoch: 1 step: 11, loss is 0.6822656989097595

2022-11-23 12:04:53 epoch: 1 step: 12, loss is 0.6779125928878784

2022-11-23 12:05:01 epoch: 1 step: 13, loss is 0.6717594265937805

2022-11-23 12:05:09 epoch: 1 step: 14, loss is 0.6642793416976929

2022-11-23 12:05:16 epoch: 1 step: 15, loss is 0.6606000065803528

2022-11-23 12:05:24 epoch: 1 step: 16, loss is 0.6534802317619324

2022-11-23 12:05:31 epoch: 1 step: 17, loss is 0.646770179271698

...

2022-11-23 12:40:33 epoch: 5 step: 51, loss is 0.46508893370628357

2022-11-23 12:40:44 epoch: 5 step: 52, loss is 0.46591484546661377

2022-11-23 12:40:53 epoch: 5 step: 53, loss is 0.46436214447021484

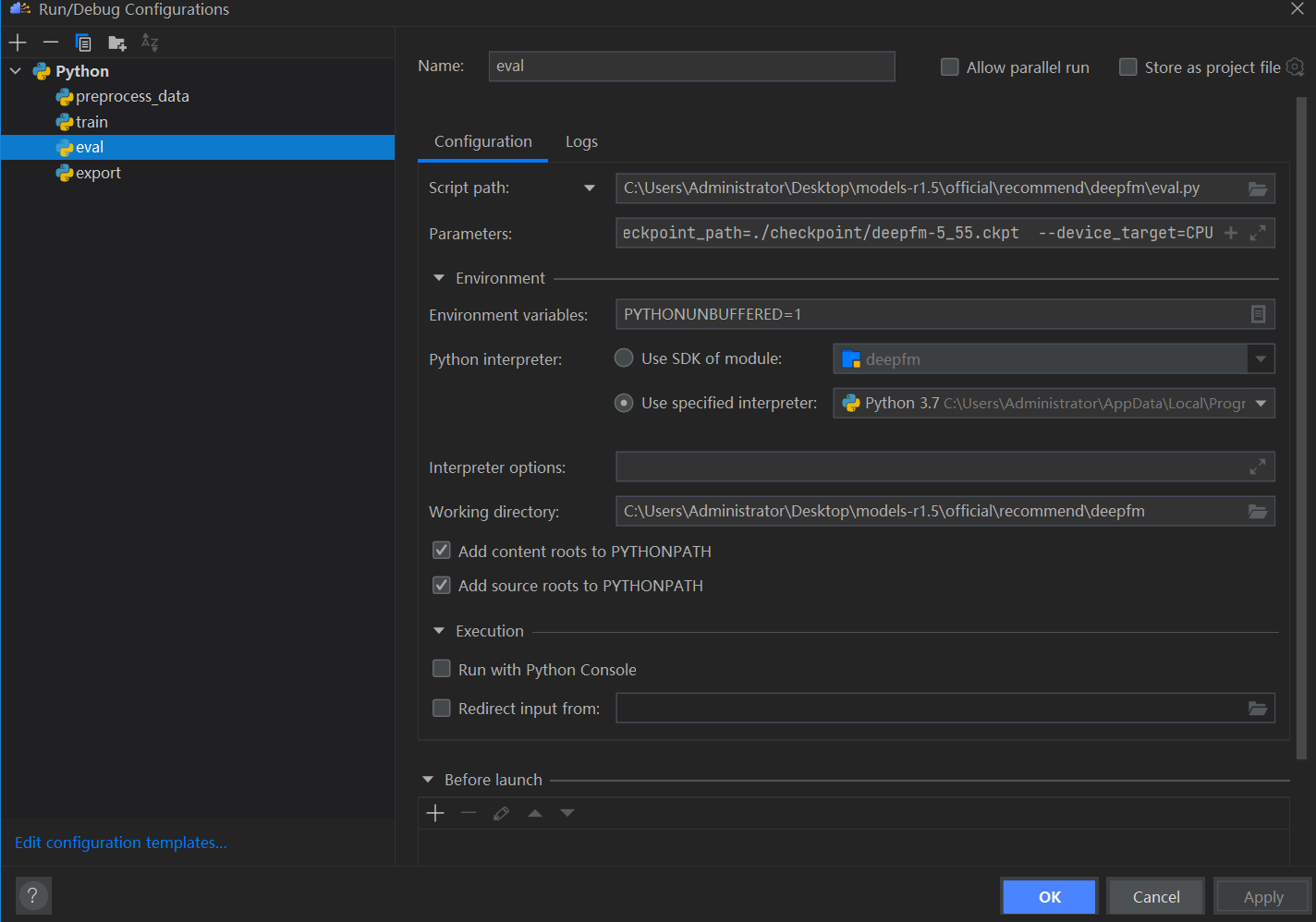
2022-11-23 12:41:00 epoch: 5 step: 54, loss is 0.47739702463150024

2022-11-23 12:41:07 epoch: 5 step: 55, loss is 0.4688734710216522

1. 评估过程

步骤 1打开deepfm\eval.py脚本文件

步骤 2设置eval.py脚本参数



步骤 3运行eval.py脚本文件

{'enable\_modelarts': 'Whether training on modelarts, default: False', 'data\_url': 'Dataset url for obs', 'train\_url': 'Training output url for obs', 'data\_path': 'Dataset path for local', 'output\_path': 'Training output path for local', 'device\_target': 'device target, support Ascend, GPU and CPU.', 'dataset\_path': 'Dataset path', 'batch\_size': 'batch size', 'ckpt\_path': 'Checkpoint path', 'eval\_file\_name': 'Auc log file path. Default: "./auc.log"', 'loss\_file\_name': 'Loss log file path. Default: "./loss.log"', 'do\_eval': 'Do evaluation or not, only support "True" or "False". Default: "True"', 'checkpoint\_path': 'Checkpoint file path', 'device\_id': 'Device id', 'ckpt\_file': 'Checkpoint file path.', 'file\_name': 'output file name.', 'file\_format': 'file format', 'result\_path': 'Result path', 'label\_path': 'label path', 'dense\_dim': 'The number of your continues fields', 'slot\_dim': 'The number of your sparse fields, it can also be called catelogy features.', 'threshold': 'Word frequency below this will be regarded as OOV. It aims to reduce the vocab size', 'train\_line\_count': 'The number of examples in your dataset', 'skip\_id\_convert': 'Skip the id convert, regarding the original id as the final id.'}

{'batch\_size': 16000,

 'checkpoint\_path': './checkpoint/deepfm-5\_55.ckpt',

 'checkpoint\_url': '',

 'ckpt\_file': '/cache/train/deepfm-5\_2582.ckpt',

 'ckpt\_file\_name\_prefix': 'deepfm',

 'ckpt\_path': '/cache/train',

 'config\_path': 'C:\\Users\\Administrator\\Desktop\\models-r1.5\\official\\recommend\\deepfm\\src\\model\_utils\\../../default\_config.yaml',

 'convert\_dtype': True,

 'data\_emb\_dim': 80,

 'data\_field\_size': 39,

 'data\_format': 1,

 'data\_path': '/cache/data',

 'data\_url': '',

 'data\_vocab\_size': 184965,

 'dataset\_path': './data/mindrecord/',

 'deep\_layer\_args': [[1024, 512, 256, 128], 'relu'],

 'dense\_dim': 13,

 'device\_id': 0,

 'device\_target': 'CPU',

 'do\_eval': 'True',

 'enable\_modelarts': False,

 'enable\_profiling': False,

 'epsilon': 5e-08,

 'eval\_callback': True,

 'eval\_file\_name': './auc.log',

 'file\_format': 'AIR',

 'file\_name': 'deepfm',

 'init\_args': [-0.01, 0.01],

 'keep\_checkpoint\_max': 50,

 'keep\_prob': 0.9,

 'l2\_coef': 8e-05,

 'label\_path': '',

 'learning\_rate': 0.0005,

 'load\_path': '/cache/checkpoint\_path',

 'loss\_callback': True,

 'loss\_file\_name': './loss.log',

 'loss\_scale': 1024.0,

 'output\_path': '/cache/train',

 'result\_path': './preprocess\_Result',

 'save\_checkpoint': True,

 'save\_checkpoint\_steps': 1,

 'skip\_id\_convert': 0,

 'slot\_dim': 26,

 'test\_num\_of\_parts': 3,

 'threshold': 100,

 'train\_epochs': 5,

 'train\_line\_count': 45840617,

 'train\_num\_of\_parts': 21,

 'train\_url': '',

 'weight\_bias\_init': ['normal', 'normal']}

Please check the above information for the configurations

[WARNING] ME(12968:9304,MainProcess):2022-11-23-12:51:59.519.86 [mindspore\dataset\engine\datasets.py:3619] WARN: global shuffle is not used.

[WARNING] ME(12968:9304,MainProcess):2022-11-23-12:52:01.458.894 [mindspore\train\model.py:893] CPU cannot support dataset sink mode currently.So the evaluating process will be performed with dataset non-sink mode.

2022-11-23 12:52:21 AUC: 0.7730511889787368, eval time: 19.70049023628235s.

Process finished with exit code 0

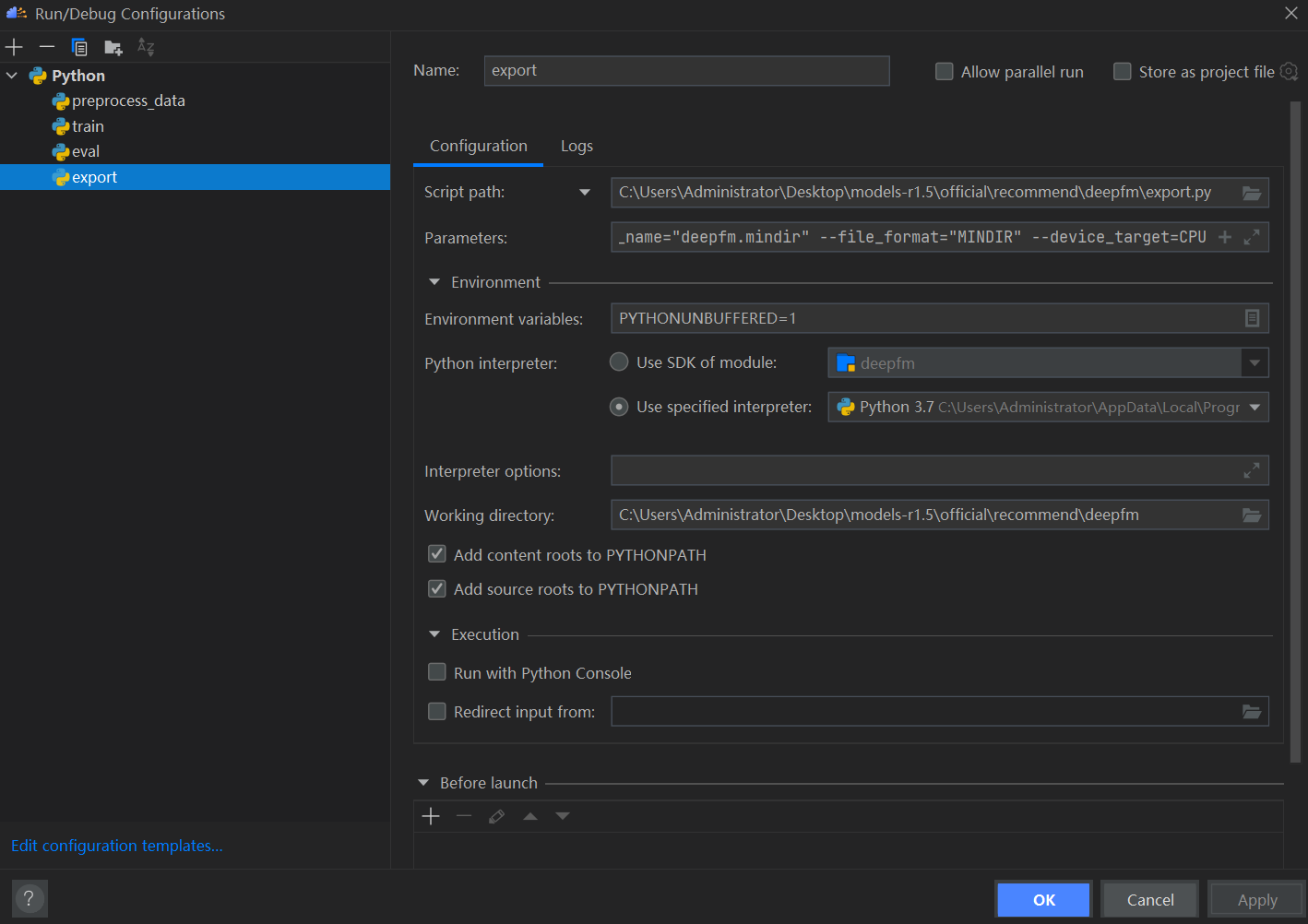
步骤 4查看auc.log日志

2022-11-23 12:52:21 AUC: 0.7730511889787368, eval time: 19.70049023628235s.

1. 2.6.5 导出MindIR

步骤 1打开deepfm\export.py脚本文件

步骤 2设置export.py脚本参数



步骤 3运行export.py脚本文件

{'enable\_modelarts': 'Whether training on modelarts, default: False', 'data\_url': 'Dataset url for obs', 'train\_url': 'Training output url for obs', 'data\_path': 'Dataset path for local', 'output\_path': 'Training output path for local', 'device\_target': 'device target, support Ascend, GPU and CPU.', 'dataset\_path': 'Dataset path', 'batch\_size': 'batch size', 'ckpt\_path': 'Checkpoint path', 'eval\_file\_name': 'Auc log file path. Default: "./auc.log"', 'loss\_file\_name': 'Loss log file path. Default: "./loss.log"', 'do\_eval': 'Do evaluation or not, only support "True" or "False". Default: "True"', 'checkpoint\_path': 'Checkpoint file path', 'device\_id': 'Device id', 'ckpt\_file': 'Checkpoint file path.', 'file\_name': 'output file name.', 'file\_format': 'file format', 'result\_path': 'Result path', 'label\_path': 'label path', 'dense\_dim': 'The number of your continues fields', 'slot\_dim': 'The number of your sparse fields, it can also be called catelogy features.', 'threshold': 'Word frequency below this will be regarded as OOV. It aims to reduce the vocab size', 'train\_line\_count': 'The number of examples in your dataset', 'skip\_id\_convert': 'Skip the id convert, regarding the original id as the final id.'}

{'batch\_size': 16000,

 'checkpoint\_path': '/cache/train/deepfm-5\_2582.ckpt',

 'checkpoint\_url': '',

 'ckpt\_file': '.\\checkpoint\\deepfm-5\_55.ckpt',

 'ckpt\_file\_name\_prefix': 'deepfm',

 'ckpt\_path': '/cache/train',

 'config\_path': 'C:\\Users\\Administrator\\Desktop\\models-r1.5\\official\\recommend\\deepfm\\src\\model\_utils\\../../default\_config.yaml',

 'convert\_dtype': True,

 'data\_emb\_dim': 80,

 'data\_field\_size': 39,

 'data\_format': 1,

 'data\_path': '/cache/data',

 'data\_url': '',

 'data\_vocab\_size': 184965,

 'dataset\_path': '/cache/data',

 'deep\_layer\_args': [[1024, 512, 256, 128], 'relu'],

 'dense\_dim': 13,

 'device\_id': 0,

 'device\_target': 'CPU',

 'do\_eval': 'True',

 'enable\_modelarts': False,

 'enable\_profiling': False,

 'epsilon': 5e-08,

 'eval\_callback': True,

 'eval\_file\_name': './auc.log',

 'file\_format': 'MINDIR',

 'file\_name': 'deepfm.mindir',

 'init\_args': [-0.01, 0.01],

 'keep\_checkpoint\_max': 50,

 'keep\_prob': 0.9,

 'l2\_coef': 8e-05,

 'label\_path': '',

 'learning\_rate': 0.0005,

 'load\_path': '/cache/checkpoint\_path',

 'loss\_callback': True,

 'loss\_file\_name': './loss.log',

 'loss\_scale': 1024.0,

 'output\_path': '/cache/train',

 'result\_path': './preprocess\_Result',

 'save\_checkpoint': True,

 'save\_checkpoint\_steps': 1,

 'skip\_id\_convert': 0,

 'slot\_dim': 26,

 'test\_num\_of\_parts': 3,

 'threshold': 100,

 'train\_epochs': 5,

 'train\_line\_count': 45840617,

 'train\_num\_of\_parts': 21,

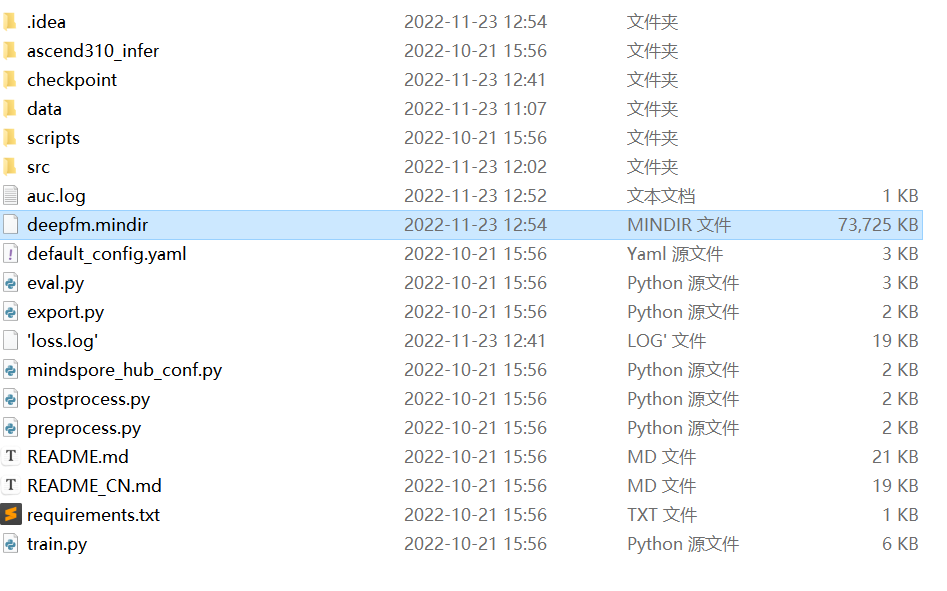
 'train\_url': '',

 'weight\_bias\_init': ['normal', 'normal']}

Please check the above information for the configurations

Process finished with exit code 0

步骤 4查看mindir模型



1. 实验总结

本章实验使用MindStudio实现了基于MindSpore框架的DeepFM广告点击率预估实验，了解了DeepFM模型的原理、训练和评估等知识点，同时也熟悉了MindSpore深度学习框架的基础用法，以及MindStudio软件的简单操作等，为后续的学习打下了稳固的基础。