Please follow below instructions to Train NER model on training/testing files:

**SCRIPT MODE**

1. Make sure all these files are present in your data directory (**train.txt, train\_box.txt, train\_image.txt, test.txt, test\_box.txt, test\_image.txt, labels.txt**)
2. Run following command to train NER model for python script based:

!python ./train.py --epochs=5 --data\_dir='./data' --output\_dir='./out' --train\_batch\_size=4 --eval\_batch\_size=4 --checkpoint\_model\_path='./output/checkpoints/'

Parameters:

* Epochs - Number of epochs to train model
* data\_dir - data directory where training and test files are present(**train.txt, train\_box.txt, train\_image.txt, test.txt, test\_box.txt, test\_image.txt, labels.txt**)
* output\_dir - output directory where all model output files are stored like model checkpoint, model metrics on test data, best model checkpoint used for inference/deployment.
* train\_batch\_size - train batch size (Adjust as per GPU memory)
* eval\_batch\_size - test batch size (Adjust as per GPU memory)
* Checkpoint\_model\_path - pass last model checkpoint files to start model training from this checkpoint

**Note:**

1. Please refer to this link for demonstration - <https://drive.google.com/drive/u/2/folders/1tyWBXTPJjSQGGN9O5Cm-5OJow4Nzorp7>
2. Open and run [**train\_layoutlm\_script.ipynb**](https://drive.google.com/drive/folders/1tyWBXTPJjSQGGN9O5Cm-5OJow4Nzorp7?usp=drive_link)notebook cells.

**NOTEBOOK TRAINING**

1. Make sure all these files are present in your data directory (**train.txt, train\_box.txt, train\_image.txt, test.txt, test\_box.txt, test\_image.txt, labels.txt**)
2. Open and run [**LayoutLMv1\_training\_script.ipynb**](https://drive.google.com/drive/u/2/folders/1nN-godvvvwOzP3MUDu0lCwjK8cVhLJ7G)notebook cells sequentially.
3. Change following directories path in notebook cells:

* Labels.txt file directory-

labels = get\_labels("/content/drive/My Drive/aws\_annotation/data/labels.txt")

[https://colab.research.google.com/drive/1sn0xCIbuOEz6GgWo3lnhaMY-VNWHW-o8?authuser=2#scrollTo=PIjINVaNRl8J line=14&uniqifier=1](https://colab.research.google.com/drive/1sn0xCIbuOEz6GgWo3lnhaMY-VNWHW-o8?authuser=2#scrollTo=PIjINVaNRl8J&line=14&uniqifier=1)

* Data directory -

args = {'local\_rank': -1,

'overwrite\_cache': True,

'data\_dir': '/content/drive/MyDrive/aws\_annotation/data',

'model\_name\_or\_path':'microsoft/layoutlm-base-uncased',

'max\_seq\_length': 512,

'model\_type': 'layoutlm',}

<https://colab.research.google.com/drive/1sn0xCIbuOEz6GgWo3lnhaMY-VNWHW-o8?authuser=2#scrollTo=9cBBA_ws0rTZ&line=5&uniqifier=1>

* Model checkpoints and output directory -

best\_model = f'models/checkpointLM1\_epoch{epoch}.pt'

for ckpt in os.listdir('models'):

if 'checkpointLM1\_epoch' in ckpt:

os.remove(f'models/{ckpt}')

if not os.path.exists(f'results/v1'):

os.mkdir(f'results/v1')

df.to\_csv(f'results/v1/log\_v1\_FUNSD\_{test\_mode}\_run{run}.csv', index = False)

<https://colab.research.google.com/drive/1sn0xCIbuOEz6GgWo3lnhaMY-VNWHW-o8?authuser=2#scrollTo=JXsd7u37jPud&line=166&uniqifier=1>

**Note:** Please refer to this link for demonstration-

<https://drive.google.com/drive/u/2/folders/1ykGJ3fD29gJYMgkZIsBYMp4shtk28YKc>