1. Refer to the README from <https://github.com/openvinotoolkit/nncf/tree/v1.7.1/third_party_integration/huggingface_transformers> to integrate NNCF into Huggingface transformers
   1. Apply the 0001-Modifications-for-NNCF-usage.patch gitpatch avail at this link to the huggingface transformers repository checked out at commit id: b0892fa0e8df02d683e05e625b3903209bff362d.
   2. Install the transformers library and the example scripts from this patched repository as described in the documentation for the huggingface transformers repository.
   3. NNCF configuration file (nncf\_mobilebert\_config\_squad\_int8.json) that was used to obtain the quantized MobileBERT for MLPerf is available in the gitpatch.
2. Here is the command line that was used for obtaining the INT8 MobileBERT. It may or may not work on the WG member’s exact HW setup. Pretrained FP32 MobileBERT FP32 backbone points to /home/user/mobilebert\_squad\_fp32 in the command; Either the member pretrain or may take the publicly available MobileBERT FP32 SQUAD baseline to specify as --*model\_name\_or\_path*

*export CUDA\_VISIBLE\_DEVICES=0,1,2; python -m torch.distributed.launch --nproc\_per\_node=3 /home/user/transformers/examples/question-answering/run\_squad.py --model\_type mobilebert --model\_name\_or\_path /home/vshampor/mobilebert\_squad\_fp32 --do\_lower\_case --do\_train --do\_eval --doc\_stride 128 --learning\_rate 3e-5 --max\_answer\_length=30 --max\_query\_length=64 --max\_seq\_length 384 --n\_best\_size=20 --num\_train\_epochs 5 --output\_dir ../models/mobilebert\_squad\_int8 --predict\_file /home/user/squad/dev-v1.1.json --per\_gpu\_train\_batch\_size=32 --per\_gpu\_eval\_batch\_size=8 --train\_file /home/user/squad/train-v1.1.json --logging\_steps=693 --save\_steps=693 --evaluate\_during\_training --nncf\_config nncf\_mobilebert\_config\_squad\_int8.json*

1. Since model training is rather non-deterministic by default, the resulting metrics of the compressed and fine-tuned model may differ slightly from what is available in the ONNX version we submitted. Also note that since NNCF is a living product, the quantization algorithm has probably changed (for better) since the date we prepared the INT8 MobileBERT, so for best reproducibility it might be prudent to take the historical NNCF version we used then (which is around [this](https://github.com/openvinotoolkit/nncf/tree/be622d4a6011e328e04c2d40779297fae4e43c45/third_party_integration/huggingface_transformers) commit) rather than the latest v1.7.1 NNCF version that was linked above. In both cases, though, the corresponding HF transformers repo commit to apply the NNCF-enabling patch to be listed is in the README file.