Get the code

1. git clone https://github.com/wuhuikx/mkl-dnn.git
2. Cd mkl-dnn
3. Git checkout vnni
4. cd scripts/
5. ./prepare_mkl.sh

Commit number

Commit number before fuse: ae00102be506ed0fe2099c6557df2aa88ad57ec1
Commit number before fuse: 3d0ad7f375aa663b36877e8e35dbeaec217c6893

Test cases

Conv3x3: (batch=2, input=(32, 258, 258), output=(64, 256, 256), kernel=(3, 3))
Conv1x1: (batch=2, input=(64, 256, 256), output=(96, 256, 256), kernel=(1, 1))

Main code files

src/cpu/jit_avx512_core_u8s8s32x_convolution.cpp
tests/gtests/test_convolution_relu_forward_common.hpp

performance

"The submit time"
"The mean time" indicate the mean time of middle 80 submits

Use the following method for testing：

- Comment the reference computation code
- Apply cache flush before and after submit

- **Conv3x3ReluConv1x1Relu** 4 op fuse
  - remove comment in line32 in cmake file
    vim cmake/platform.cmake +32
    31 #add_definitions(-DNON_FUSE)
    32 add_definitions(-DCONV11_FUSE)

  - vim run_skx.sh
    taskset -c 0-27 numactl -l ./build/tests/gtests/test_convolution_relu_forward_u8s8s32

  - bash ./build.sh
    bash ./run_skx.sh

- **Conv3x3Relu + Conv1x1Relu 2 op fuse**
  - remove comment in line31 in cmake file
    vim cmake/platform.cmake +31
    31 add_definitions(-DNON_FUSE)
    32 #add_definitions(-DCONV11_FUSE)

- vim run_skx.sh
  taskset -c 0-27 numactl -l ./build/tests/gtests/test_convolution_relu_forward_u8s8s32

- bash ./build.sh
  bash ./run_skx.sh

- **Conv3x3 + Relu + Conv1x1 + Relu non-fuse**
  - remove comment in line31 in cmake file
    vim cmake/platform.cmake +31
    31 add_definitions(-DNON_FUSE)
    32 #add_definitions(-DCONV11_FUSE)

  - vim run_skx.sh
    taskset -c 0-27 numactl -
    l ./build/tests/gtests/test_convolution_relu_forward_u8s8s32_discrete

  - bash ./build.sh
    bash ./run_skx.sh