Item 289

git_comments:

git_commits:

1. **summary:** [int8] Add MobileNetV2_1.0 & ResNet18 Quantization (#14823)
   message: [int8] Add MobileNetV2_1.0 & ResNet18 Quantization (#14823) * add resnet18 and mobilenetv2 models * add readme * support mkldnn s8s8 goihw16g weight format * fix_readme_typo

github_issues:

github_issues_comments:

github_pulls:

1. **title:** [int8] Add MobileNetV2_1.0 & ResNet18 Quantization
   body: ## Description ## Add MobileNetV2_1.0 & ResNet18 Quantization. ResNet18 Performance on Skylake 8180 28c resnet18_v1 | fp32 | int8 | speedup :--: | :--: | :--: | :--: 1 | 309.61 | 492.18 | 1.59 64 | 810.82 | 1341.55 | 1.65 accuracy | 70.07%/89.30% | 69.85%/89.23% | #14819 will improve mobilenetv2 fp32/int8 performance mobilenetv2_1.0 | fp32 | int8 | speedup | fp32_opt | int8_opt | speedup :--:|:--: |:--: | :--:|:--: |:--: |:--: 1 | 75.22 | 162.12 | 2.16 | 240.51 | 413.92 | 1.72 64 | 291.63 | 469.28 | 1.61 | 795.86 | 3137.77 | 3.94 accuracy | 70.14%/89.60% | 63.62%/84.84% | | 70.14%/89.60% | 69.53%/89.24% | @pengzhao-intel @ZhennanQin ## Checklist ## ### Essentials ### Please feel free to remove inapplicable items for your PR. - [ ] The PR title starts with [MXNET-$JIRA_ID], where $JIRA_ID refers to the relevant [JIRA issue](https://issues.apache.org/jira/projects/MXNET/issues) created (except PRs with tiny changes) - [ ] Changes are complete (i.e. I finished coding on this PR) - [ ] All changes have test coverage: - Unit tests are added for small changes to verify correctness (e.g. adding a new operator) - Nightly tests are added for complicated/long-running ones (e.g. changing distributed kvstore) - Build tests will be added for build configuration changes (e.g. adding a new build option with NCCL) - [ ] Code is well-documented: - For user-facing API changes, API doc string has been updated. - For new C++ functions in header files, their functionalities and arguments are documented. - For new examples, README.md is added to explain the what the example does, the source of the dataset, expected performance on test set and reference to the original paper if applicable - Check the API doc at http://mxnet-ci-doc.s3-accelerate.dualstack.amazonaws.com/PR-$PR_ID/$BUILD_ID/index.html - [ ] To the my best knowledge, examples are either not affected by this change, or have been fixed to be compatible with this change ### Changes ### - [ ] Feature1, tests, (and when applicable, API doc) - [ ] Feature2, tests, (and when applicable, API doc) ## Comments ## - If this change is a backward incompatible change, why must this change be made. - Interesting edge cases to note here

github_pulls_comments:

1. cc @zhreshold
2. @mxnet-label-bot add [Quantization, Example]
3. Merging now. @xinyu-intel will refactor the script in the next PR.

github_pulls_reviews:

1. What's this? It's first time for us to have this format?
2. yes, when quantize s8s8 group conv.
3. is there any exception that the rgb_mean dna std is not the same? otherwise repeatively coding it looks redundant
4. Agree with @zhreshold, along with enabling more models, we don't need to show how to reproduce each one since most of the command is very similar. We need to define a template for the user to reproduce any existed models for classification networks.
5. @zhreshold @pengzhao-intel agree, I'll refactor this script along with enabling more models next time.

jira_issues:

jira_issues_comments: