Quantization API Reference

torch.quantization

This module contains Eager mode quantization APIs.

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 9)**
>
> Unknown directive type "currentmodule".
>
> ```
> .. currentmodule:: torch.quantization
> ```

Top level APIs

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 14)**
>
> Unknown directive type "autosummary".
>
> ```
> .. autosummary::
>     :toctree: generated
>     :nosignatures:
>     :template: classtemplate.rst
>
>     quantize
>     quantize_dynamic
>     quantize_qat
>     prepare
>     prepare_qat
>     convert
> ```

Preparing model for quantization

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 29)**
>
> Unknown directive type "autosummary".
>
> ```
> .. autosummary::
>     :toctree: generated
>     :nosignatures:
>     :template: classtemplate.rst
>
>     fuse_modules
>     QuantStub
>     DeQuantStub
>     QuantWrapper
>     add_quant_dequant
> ```

Utility functions

> **System Message: ERROR/3** (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 43)**
>
> Unknown directive type "autosummary".
>
> ```
> .. autosummary::
>     :toctree: generated
>     :nosignatures:
>     :template: classtemplate.rst
>
>     add_observer_
>     swap_module
>     propagate_qconfig_
>     default_eval_fn
>     get_observer_dict
> ```

torch.quantization.quantize_fx

This module contains FX graph mode quantization APIs (prototype).

> **System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 59`)**
>
> Unknown directive type "currentmodule".
>
> ```
> .. currentmodule:: torch.quantization.quantize_fx
> ```

> **System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 61`)**
>
> Unknown directive type "autosummary".
>
> ```
> .. autosummary::
>     :toctree: generated
>     :nosignatures:
>     :template: classtemplate.rst
>
>     prepare_fx
>     prepare_qat_fx
>     convert_fx
>     fuse_fx
> ```

torch (quantization related functions)

This describes the quantization related functions of the *torch* namespace.

> **System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 76`)**
>
> Unknown directive type "currentmodule".
>
> ```
> .. currentmodule:: torch
> ```

> **System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 78`)**
>
> Unknown directive type "autosummary".
>
> ```
> .. autosummary::
>     :toctree: generated
>     :nosignatures:
>     :template: classtemplate.rst
>
>     quantize_per_tensor
>     quantize_per_channel
>     dequantize
> ```

torch.Tensor (quantization related methods)

Quantized Tensors support a limited subset of data manipulation methods of the regular full-precision tensor.

> **System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 93`)**
>
> Unknown directive type "currentmodule".
>
> ```
> .. currentmodule:: torch.Tensor
> ```

> **System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 95`)**
>
> Unknown directive type "autosummary".

```
.. autosummary::
   :toctree: generated
   :nosignatures:
   :template: classtemplate.rst

   view
   as_strided
   expand
   flatten
   select
   ne
   eq
   ge
   le
   gt
   lt
   copy_
   clone
   dequantize
   equal
   int_repr
   max
   mean
   min
   q_scale
   q_zero_point
   q_per_channel_scales
   q_per_channel_zero_points
   q_per_channel_axis
   resize_
   sort
   topk
```

torch.quantization.observer

This module contains observers which are used to collect statistics about the values observed during calibration (PTQ) or training (QAT).

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, line 135)

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.quantization.observer
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, line 137)

Unknown directive type "autosummary".

```
.. autosummary::
   :toctree: generated
   :nosignatures:
   :template: classtemplate.rst

   ObserverBase
   MinMaxObserver
   MovingAverageMinMaxObserver
   PerChannelMinMaxObserver
   MovingAveragePerChannelMinMaxObserver
   HistogramObserver
   PlaceholderObserver
   RecordingObserver
   NoopObserver
   get_observer_state_dict
   load_observer_state_dict
   default_observer
   default_placeholder_observer
   default_debug_observer
   default_weight_observer
   default_histogram_observer
   default_per_channel_weight_observer
   default_dynamic_quant_observer
   default_float_qparams_observer
```

torch.quantization.fake_quantize

This module implements modules which are used to perform fake quantization during QAT.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 168`)

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.quantization.fake_quantize
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 170`)

Unknown directive type "autosummary".

```
.. autosummary::
    :toctree: generated
    :nosignatures:
    :template: classtemplate.rst

    FakeQuantizeBase
    FakeQuantize
    FixedQParamsFakeQuantize
    FusedMovingAvgObsFakeQuantize
    default_fake_quant
    default_weight_fake_quant
    default_per_channel_weight_fake_quant
    default_histogram_fake_quant
    default_fused_act_fake_quant
    default_fused_wt_fake_quant
    default_fused_per_channel_wt_fake_quant
    disable_fake_quant
    enable_fake_quant
    disable_observer
    enable_observer
```

torch.quantization.qconfig

This module defines *QConfig* objects which are used to configure quantization settings for individual ops.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 197`)

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.quantization.qconfig
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 199`)

Unknown directive type "autosummary".

```
.. autosummary::
    :toctree: generated
    :nosignatures:
    :template: classtemplate.rst

    QConfig
    default_qconfig
    default_debug_qconfig
    default_per_channel_qconfig
    default_dynamic_qconfig
    float16_dynamic_qconfig
    float16_static_qconfig
    per_channel_dynamic_qconfig
    float_qparams_weight_only_qconfig
    default_qat_qconfig
    default_weight_only_qconfig
    default_activation_only_qconfig
    default_qat_qconfig_v2
```

torch.nn.intrinsic

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 220`)

Unknown directive type "automodule".

```
.. automodule:: torch.nn.intrinsic
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 221`)

Unknown directive type "automodule".

```
.. automodule:: torch.nn.intrinsic.modules
```

This module implements the combined (fused) modules conv + relu which can then be quantized.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 226`)

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.nn.intrinsic
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 228`)

Unknown directive type "autosummary".

```
.. autosummary::
    :toctree: generated
    :nosignatures:
    :template: classtemplate.rst

    ConvReLU1d
    ConvReLU2d
    ConvReLU3d
    LinearReLU
    ConvBn1d
    ConvBn2d
    ConvBn3d
    ConvBnReLU1d
    ConvBnReLU2d
    ConvBnReLU3d
    BNReLU2d
    BNReLU3d
```

torch.nn.intrinsic.qat

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 248`)

Unknown directive type "automodule".

```
.. automodule:: torch.nn.intrinsic.qat
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, line 249`)

Unknown directive type "automodule".

```
.. automodule:: torch.nn.intrinsic.qat.modules
```

This module implements the versions of those fused operations needed for quantization aware training.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst**, line 255)**

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.nn.intrinsic.qat
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst**, line 257)**

Unknown directive type "autosummary".

```
.. autosummary::
   :toctree: generated
   :nosignatures:
   :template: classtemplate.rst

   LinearReLU
   ConvBn1d
   ConvBnReLU1d
   ConvBn2d
   ConvBnReLU2d
   ConvReLU2d
   ConvBn3d
   ConvBnReLU3d
   ConvReLU3d
   update_bn_stats
   freeze_bn_stats
```

torch.nn.intrinsic.quantized

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst**, line 276)**

Unknown directive type "automodule".

```
.. automodule:: torch.nn.intrinsic.quantized
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst**, line 277)**

Unknown directive type "automodule".

```
.. automodule:: torch.nn.intrinsic.quantized.modules
```

This module implements the quantized implementations of fused operations like conv + relu. No BatchNorm variants as it's usually folded into convolution for inference.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst**, line 284)**

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.nn.intrinsic.quantized
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst**, line 286)**

Unknown directive type "autosummary".

```
.. autosummary::
   :toctree: generated
   :nosignatures:
   :template: classtemplate.rst

   BNReLU2d
   BNReLU3d
   ConvReLU1d
```

```
    ConvReLU2d
    ConvReLU3d
    LinearReLU
```

torch.nn.intrinsic.quantized.dynamic

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, **line 300**)

Unknown directive type "automodule".

```
.. automodule:: torch.nn.intrinsic.quantized.dynamic
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, **line 301**)

Unknown directive type "automodule".

```
.. automodule:: torch.nn.intrinsic.quantized.dynamic.modules
```

This module implements the quantized dynamic implementations of fused operations like linear + relu.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, **line 306**)

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.nn.intrinsic.quantized.dynamic
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, **line 308**)

Unknown directive type "autosummary".

```
.. autosummary::
    :toctree: generated
    :nosignatures:
    :template: classtemplate.rst

    LinearReLU
```

torch.nn.qat

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, **line 317**)

Unknown directive type "automodule".

```
.. automodule:: torch.nn.qat
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, **line 318**)

Unknown directive type "automodule".

```
.. automodule:: torch.nn.qat.modules
```

This module implements versions of the key nn modules **Conv2d()** and **Linear()** which run in FP32 but with rounding applied to simulate the effect of INT8 quantization.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, **line 324**)

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.nn.qat
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 326)**

Unknown directive type "autosummary".

```
.. autosummary::
   :toctree: generated
   :nosignatures:
   :template: classtemplate.rst

   Conv2d
   Conv3d
   Linear
```

torch.nn.qat.dynamic

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 337)**

Unknown directive type "automodule".

```
.. automodule:: torch.nn.qat.dynamic
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 338)**

Unknown directive type "automodule".

```
.. automodule:: torch.nn.qat.dynamic.modules
```

This module implements versions of the key nn modules such as **Linear()** which run in FP32 but with rounding applied to simulate the effect of INT8 quantization and will be dynamically quantized during inference.

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 344)**

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.nn.qat.dynamic
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 346)**

Unknown directive type "autosummary".

```
.. autosummary::
   :toctree: generated
   :nosignatures:
   :template: classtemplate.rst

   Linear
```

torch.nn.quantized

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst, **line 355)**

Unknown directive type "automodule".

```
.. automodule:: torch.nn.quantized
```

System Message: ERROR/3 (D:\onboarding-resources\sample-onboarding-resources\pytorch-

This module implements the quantized versions of the nn layers such as ~`torch.nn.Conv2d` and *torch.nn.ReLU*.

torch.nn.quantized.functional

This module implements the quantized versions of the functional layers such as ~`torch.nn.functional.conv2d` and *torch.nn.functional.relu*. Note: :meth:`~torch.nn.functional.relu` supports quantized inputs.

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.nn.quantized.functional
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 403)**

Unknown directive type "autosummary".

```
.. autosummary::
    :toctree: generated
    :nosignatures:
    :template: classtemplate.rst

    avg_pool2d
    avg_pool3d
    adaptive_avg_pool2d
    adaptive_avg_pool3d
    conv1d
    conv2d
    conv3d
    interpolate
    linear
    max_pool1d
    max_pool2d
    celu
    leaky_relu
    hardtanh
    hardswish
    threshold
    elu
    hardsigmoid
    clamp
    upsample
    upsample_bilinear
    upsample_nearest
```

torch.nn.quantized.dynamic

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 433)**

Unknown directive type "automodule".

```
.. automodule:: torch.nn.quantized.dynamic
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 434)**

Unknown directive type "automodule".

```
.. automodule:: torch.nn.quantized.dynamic.modules
```

Dynamically quantized :class:`~torch.nn.Linear`, :class:`~torch.nn.LSTM`, :class:`~torch.nn.LSTMCell`, :class:`~torch.nn.GRUCell`, and :class:`~torch.nn.RNNCell`.

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 436); *backlink***

Unknown interpreted text role "class".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 436); *backlink***

Unknown interpreted text role "class".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 436);** *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 436);** *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 436);** *backlink*

Unknown interpreted text role "class".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 440)**

Unknown directive type "currentmodule".

```
.. currentmodule:: torch.nn.quantized.dynamic
```

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 442)**

Unknown directive type "autosummary".

```
.. autosummary::
   :toctree: generated
   :nosignatures:
   :template: classtemplate.rst

   Linear
   LSTM
   GRU
   RNNCell
   LSTMCell
   GRUCell
```

Quantized dtypes and quantization schemes

Note that operator implementations currently only support per channel quantization for weights of the **conv** and **linear** operators. Furthermore, the input data is mapped linearly to the the quantized data and vice versa as follows:

Quantization:

$$Q_{out} = \text{clamp}(x_{input} / s + z, Q_{min}, Q_{max})$$

Dequantization:

$$x_{out} = (Q_{input} - z) * s$$

where clamp(.) is the same as :func:`~torch.clamp` while the scale s and zero point z are then computed as decribed in :class:`~torch.ao.quantization.observer.MinMaxObserver`, specifically:

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 472);** *backlink*

Unknown interpreted text role "func".

System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`**, line 472);** *backlink*

Unknown interpreted text role "class".

if Symmetric:

$$s = 2\max(|x_{\min}|, x_{\max}) / (Q_{\max} - Q_{\min})$$

$$z = \begin{cases} 0 & \text{if dtype is qint8} \\ 128 & \text{otherwise} \end{cases}$$

Otherwise:

$$s = (x_{\max} - x_{\min}) / (Q_{\max} - Q_{\min})$$

$$z = Q_{\min} - \text{round}(x_{\min} / s)$$

where $[x_{\min}, x_{\max}]$ denotes the range of the input data while $Q_{\min}$ and $Q_{\max}$ are respectively the minimum and maximum values of the quantized dtype.

Note that the choice of s and z implies that zero is represented with no quantization error whenever zero is within the range of the input data or symmetric quantization is being used.

Additional data types and quantization schemes can be implemented through the custom operator mechanism.

- :attr:`torch.qscheme` â€" Type to describe the quantization scheme of a tensor. Supported types:

  System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, line 501); *backlink*

  Unknown interpreted text role "attr".

  - :attr:`torch.per_tensor_affine` â€" per tensor, asymmetric

    System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, line 504); *backlink*

    Unknown interpreted text role "attr".

  - :attr:`torch.per_channel_affine` â€" per channel, asymmetric

    System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, line 505); *backlink*

    Unknown interpreted text role "attr".

  - :attr:`torch.per_tensor_symmetric` â€" per tensor, symmetric

    System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, line 506); *backlink*

    Unknown interpreted text role "attr".

  - :attr:`torch.per_channel_symmetric` â€" per channel, symmetric

    System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-resources\pytorch-master\docs\source\[pytorch-master][docs][source]quantization-support.rst`, line 507); *backlink*

    Unknown interpreted text role "attr".

- `torch.dtype` â€" Type to describe the data. Supported types:

  - :attr:`torch.quint8` â€" 8-bit unsigned integer

    System Message: ERROR/3 (`D:\onboarding-resources\sample-onboarding-`

- :attr:`torch.qint8` â€" 8-bit signed integer

- :attr:`torch.qint32` â€" 32-bit signed integer