**Help Document**

**1. Quantization Methods:**

Here are the quantization methods supported by the code:

| Quantization Method | Description |
|---|---|
| **not_quantized** | Recommended. Fast conversion. Slow inference, big files. |
| **fast_quantized** | Recommended. Fast conversion. OK inference, OK file size. |
| **quantized** | Recommended. Slow conversion. Fast inference, small files. |
| **f32** | Not recommended. Retains 100% accuracy, but super slow and memory hungry. |
| **f16** | Fastest conversion + retains 100% accuracy. Slow and memory hungry. |
| **q8_0** | Fast conversion. High resource use, but generally acceptable. |
| **q4_k_m** | Recommended. Uses Q6_K for half of the attention.wv and feed_forward.w2 tensors, else Q4_K. |
| **q5_k_m** | Recommended. Uses Q6_K for half of the attention.wv and feed_forward.w2 tensors, else Q5_K. |
| **q2_k** | Uses Q4_K for the attention.vw and feed_forward.w2 tensors, Q2_K for the other tensors. |
| **q3_k_l** | Uses Q5_K for the attention.wv, attention.wo, and feed_forward.w2 tensors, else Q3_K. |
| **q3_k_m** | Uses Q4_K for the attention.wv, attention.wo, and feed_forward.w2 tensors, else Q3_K. |
| **q3_k_s** | Uses Q3_K for all tensors. |
| **q4_0** | Original quant method, 4-bit. |
| **q4_1** | Higher accuracy than q4_0 but not as high as q5_0. However, it has quicker inference than q5 models. |
| **q5_0** | Higher accuracy, higher resource usage and slower inference. |
| **q5_1** | Even higher accuracy, resource usage and slower inference. |

| Quantization Method | Description |
|---|---|
| **q6_k** | Uses Q8_K for all tensors. |

---

**2. Pre-Quantized Models:**

Here is the list of 4-bit pre-quantized models that can be used for faster downloading and avoiding out-of-memory errors (OOMs):

- **unsloth/Meta-Llama-3.1-8B-bnb-4bit**
- **unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit**
- **unsloth/Meta-Llama-3.1-70B-bnb-4bit**
- **unsloth/Meta-Llama-3.1-405B-bnb-4bit**
- **unsloth/Mistral-Nemo-Base-2407-bnb-4bit**
- **unsloth/Mistral-Nemo-Instruct-2407-bnb-4bit**
- **unsloth/mistral-7b-v0.3-bnb-4bit**
- **unsloth/mistral-7b-instruct-v0.3-bnb-4bit**
- **unsloth/Phi-3.5-mini-instruct**
- **unsloth/Phi-3-medium-4k-instruct**
- **unsloth/gemma-2-9b-bnb-4bit**
- **unsloth/gemma-2-27b-bnb-4bit**

For more models, you can visit the Unsloth Hugging Face repository.

---

**3. How to Set Parameters:**

**Model Settings:**

- You can select from the pre-quantized models or other available models on the Hugging Face platform.
- Example: model_name = "unsloth/Phi-3-mini-4k-instruct"

**Hugging Face Token:**

- Make sure you have a valid Hugging Face token.
- Example: hf_token = "your_huggingface_token"

**Dataset Settings:**

- Set the dataset repository from Hugging Face.
- Example: dataset_repo = "nimesh7814/NBRO-Chatbot-V1"

**Sequence Length:**

- The maximum sequence length must be set based on the model's capability. For example:
    - max_seq_length = 2048

**Quantization Settings:**

- To apply quantization, choose one or multiple methods from the allowed quantization list.
- Example: selected_quant_methods = ["f16", "q5_k_m"]

---

**4. Training Arguments:**

You can modify key training parameters in the script. Below are the customizable settings:

| Parameter | Description | Example |
|---|---|---|
| **max_steps** | Total number of training steps | max_steps = 5000 |
| **learning_rate** | Learning rate for optimization | learning_rate = 1e-4 |
| **save_steps** | Steps interval to save the model | save_steps = 500 |
| **eval_steps** | Steps interval for evaluation | eval_steps = 100 |

---

**5. Chat with Model After Training:**

After the model finishes training, you can choose to chat with it interactively. When prompted:

- **yes**: To start a conversation with the model.
- **no**: To skip the conversation and finish the run.

---

**6. Additional Resources:**

For more detailed information and additional models, visit the Unsloth documentation.