MNIST Examples for GGML（GGML 的 MNIST 示例）

These are simple examples of how to use GGML for inferencing. The first example uses convolutional neural network (CNN), the second one uses fully connected neural network.

- 这些是如何使用 GGML 进行推理的简单示例。第一个例子使用卷积神经网络 (CNN)，第二个例子使用全连接神经网络。

MNIST with CNN（使用CNN的MNIST）

This implementation achieves ~99% accuracy on the MNIST test set.

- 这个实现在MNIST测试集上达到了大约99%的正确率

Training the model（训练模型）

Setup the Python environemt and build the examples according to the main README.

- 设置python环境
- 根据main README构建examples

Use the `mnist-cnn.py` script to train the model and convert it to GGUF format:

- 使用mnist-cnn.py脚本训练模型并且将训练后的模型转化为GGUF格式

```
$ python3 ../examples/mnist/mnist-cnn.py train mnist-cnn-model
...
Keras model saved to 'mnist-cnn-model'
```

Convert the model to GGUF format:

```
$ python3 ../examples/mnist/mnist-cnn.py convert mnist-cnn-model
...
Model converted and saved to 'mnist-cnn-model.gguf'
```

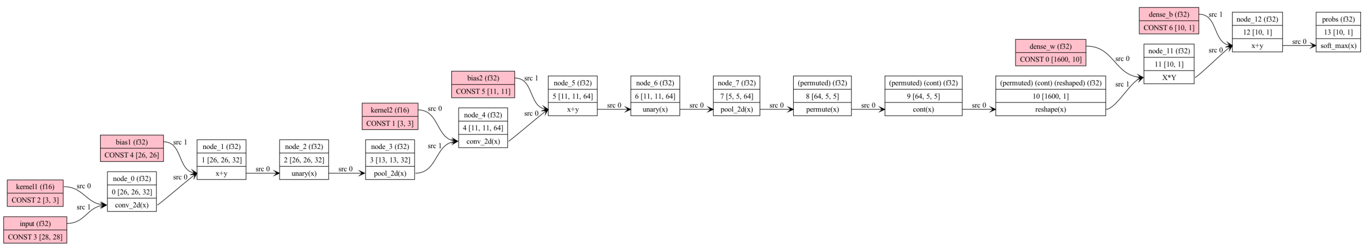
Running the example

```
$ ./bin/mnist-cnn mnist-cnn-model.gguf ../examples/mnist/models/mnist/t10k-
images.idx3-ubyte
main: loaded model in     5.17 ms
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ * * * * * _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ * * * * * * * * _ _ _ _ _ _ _ _
```

```
_ _ _ _ _ _ _ _ _ _ _ _ * * * * * _ _ _ * * _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ * * _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ * * * _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ * * * _ _ _ _ _ _ _ _ * _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ * * * _ _ _ _ _ _ _ _ * * _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ * * * _ _ _ _ _ _ _ _ * * _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ * * * _ _ _ _ * * * * * _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ * * * * * * * * * _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ * * * * * * * * * * _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ * * * * * * _ _ * * * _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ * * * _ _ _ _ _ _ _ * * * _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ * * _ _ _ _ _ _ _ _ _ * * _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ * * _ _ _ _ _ _ _ _ _ * * * _ _ _ _ _ _ _ _
_ _ _ _ _ _ * * _ _ _ _ _ _ _ _ _ * * * _ _ _ _ _ _ _ _
_ _ _ _ _ _ * * * _ _ _ _ _ _ _ _ * * * _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ * * * _ _ _ _ _ _ * * * _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ * * * * * * * * * * _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ * * * * * * _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _
_ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

ggml_graph_dump_dot: dot -Tpng mnist-cnn.dot -o mnist-cnn.dot.png && open mnist-cnn.dot.png
main: predicted digit is 8
```

Computation graph:



MNIST with fully connected network（使用全连接的MNIST）

A fully connected layer + relu, followed by a fully connected layer + softmax.

Training the Model（训练model）

A Google Colab notebook for training a simple two-layer network to recognize digits is located here. You can use this to save a pytorch model to be converted to ggml format.

- 此处有一个用于训练简单的两层网络来识别数字的 Google Colab 笔记本。您可以使用它来保存要转换为 ggml 格式的 pytorch 模型。

Colab

GGML "format" is whatever you choose for efficient loading. In our case, we just save the hyperparameters used plus the model weights and biases. Run convert-h5-to-ggml.py to convert your pytorch model. The

output format is（GGML"格式"是您选择的任何格式，以便高效加载。在我们的例子中，我们只保存使用的超参数以及模型权重和偏差。运行 convert-h5-to-ggml.py 来转换您的 pytorch 模型。输出格式为)：

- magic constant (int32)
- repeated list of tensors
- number of dimensions of tensor (int32)
- tensor dimension (int32 repeated)
- values of tensor (int32)

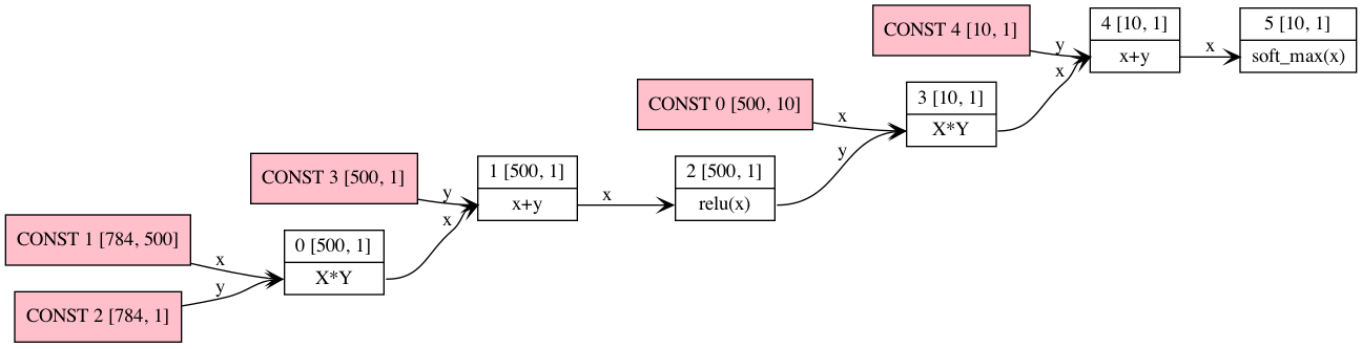
Run `convert-h5-to-ggml.py mnist_model.state_dict` where `mnist_model.state_dict` is the saved pytorch model from the Google Colab. For quickstart, it is included in the mnist/models directory.

```
mkdir -p models/mnist
python3 ../examples/mnist/convert-h5-to-ggml.py
../examples/mnist/models/mnist/mnist_model.state_dict
```

Running the example

```
./bin/mnist ./models/mnist/ggml-model-f32.bin ../examples/mnist/models/mnist/t10k-
images.idx3-ubyte
```

Computation graph:



Web demo

The example can be compiled with Emscripten like this:

```
cd examples/mnist
emcc -I../../include -I../../include/ggml -I../../examples ../../src/ggml.c
../../src/ggml-quants.c main.cpp -o web/mnist.js -s
EXPORTED_FUNCTIONS='["_wasm_eval","_wasm_random_digit","_malloc","_free"]' -s
EXPORTED_RUNTIME_METHODS='["ccall"]' -s ALLOW_MEMORY_GROWTH=1 --preload-file
models/mnist
```

Online demo: https://mnist.ggerganov.com