Supplementary code for the Build a Large Language Model From Scratch book by Sebastian Raschka

Code repository: https://github.com/rasbt/LLMs-from-scratch

Appendix A: Introduction to PyTorch (Part 2)

A.9 Optimizing training performance with GPUs

A.9.1 PyTorch computations on GPU devices

In [1]:
```
import torch

print(torch.__version__)
```

```
2.4.0+cu121
```

In [2]:
```
print(torch.cuda.is_available())
```

```
True
```

In [3]:
```
tensor_1 = torch.tensor([1., 2., 3.])
tensor_2 = torch.tensor([4., 5., 6.])

print(tensor_1 + tensor_2)
```

```
tensor([5., 7., 9.])
```

In [4]:
```
tensor_1 = tensor_1.to("cuda")
tensor_2 = tensor_2.to("cuda")

print(tensor_1 + tensor_2)
```

```
tensor([5., 7., 9.], device='cuda:0')
```

In [5]:
```
tensor_1 = tensor_1.to("cpu")
print(tensor_1 + tensor_2)
```

```
---------------------------------------------------------------------------
RuntimeError                              Traceback (most recent call last)
/tmp/ipykernel_2321/2079609735.py in <cell line: 2>()
      1 tensor_1 = tensor_1.to("cpu")
----> 2 print(tensor_1 + tensor_2)

RuntimeError: Expected all tensors to be on the same device, but found at least two devices, cuda:0 and cpu!
```

A.9.2 Single-GPU training

In [6]:
```
X_train = torch.tensor([
    [-1.2, 3.1],
    [-0.9, 2.9],
    [-0.5, 2.6],
    [2.3, -1.1],
    [2.7, -1.5]
])

y_train = torch.tensor([0, 0, 0, 1, 1])

X_test = torch.tensor([
    [-0.8, 2.8],
    [2.6, -1.6],
])

y_test = torch.tensor([0, 1])
```

In [7]:
```
from torch.utils.data import Dataset


class ToyDataset(Dataset):
    def __init__(self, X, y):
        self.features = X
        self.labels = y

    def __getitem__(self, index):
        one_x = self.features[index]
        one_y = self.labels[index]
        return one_x, one_y

    def __len__(self):
        return self.labels.shape[0]

train_ds = ToyDataset(X_train, y_train)
test_ds = ToyDataset(X_test, y_test)
```

In [8]:
```
from torch.utils.data import DataLoader

torch.manual_seed(123)

train_loader = DataLoader(
    dataset=train_ds,
    batch_size=2,
    shuffle=True,
    num_workers=0,
    drop_last=True
)

test_loader = DataLoader(
    dataset=test_ds,
    batch_size=2,
    shuffle=False,
    num_workers=0
)
```

Solution : Annotations Part 1

1. Function Class f_θ:

The **function class** f_θ refers to the neural network architecture that maps input data x to output logits. In this case, the function class is defined by the neural network model that consists of two hidden layers and an output layer. The network architecture is:

1. **Input layer**: The input data has num_inputs dimensions.
2. **1st hidden layer**: A fully connected (linear) layer with 30 neurons and a ReLU activation function.
3. **2nd hidden layer**: Another fully connected (linear) layer with 20 neurons and a ReLU activation function.
4. **Output layer**: A fully connected (linear) layer with $\text{num}\_\text{outputs}$ neurons, which outputs the logits for the final predictions.

Mathematically, the function class $f_\theta(x)$ can be represented as:

$$f_\theta(x) = W_3 \cdot \text{ReLU}(W_2 \cdot \text{ReLU}(W_1 \cdot x + b_1) + b_2) + b_3$$

where:

- W_1, b_1 are the weights and biases of the 1st hidden layer.
- W_2, b_2 are the weights and biases of the 2nd hidden layer.
- W_3, b_3 are the weights and biases of the output layer.
- ReLU is the ReLU activation function applied element-wise after the first two linear layers.

The function class describes how the input is transformed through the layers and activation functions to produce the output logits.

2. Parameter Space θ:

The **parameter space** θ refers to the set of all learnable parameters (weights and biases) in the neural network. In this model, θ includes:

- W_1 and b_1: Weights and biases of the 1st hidden layer (dimensions: [num_inputs, 30] for weights and [30] for biases).
- W_2 and b_2: Weights and biases of the 2nd hidden layer (dimensions: [30, 20] for weights and [20] for biases).
- W_3 and b_3: Weights and biases of the output layer (dimensions: [20, num_outputs] for weights and [num_outputs] for biases).

These parameters are learned during training by minimizing a loss function through optimization algorithms like stochastic gradient descent (SGD) or Adam. The parameter space θ is the collection of all these weights and biases that define the transformations applied by the neural network to the input data.

In [9]:
```
class NeuralNetwork(torch.nn.Module):
    def __init__(self, num_inputs, num_outputs):
        super().__init__()

        self.layers = torch.nn.Sequential(

            # 1st hidden layer
            torch.nn.Linear(num_inputs, 30),
            torch.nn.ReLU(),

            # 2nd hidden layer
            torch.nn.Linear(30, 20),
            torch.nn.ReLU(),

            # output layer
            torch.nn.Linear(20, num_outputs),
        )

    def forward(self, x):
        logits = self.layers(x)
        return logits
```

Solution Annotations : Part 2

1. Loss Function l:

The **loss function** used in the code is **cross-entropy loss**, which is applied to the predicted logits and the true labels:

```
loss = F.cross_entropy(logits, labels)
```

- **Cross-entropy loss** is used for classification tasks and measures the difference between the predicted class probabilities (after applying softmax) and the true class labels. It computes the negative log-likelihood of the true class based on the predicted probabilities.
- For a single data point (x_i, y_i), the cross-entropy loss $l(f_\theta(x_i), y_i)$ is given by:

$$l(f_\theta(x_i), y_i) = -\log\left(\frac{\exp(f_\theta(x_i)_{y_i})}{\sum_{k=1}^{K} \exp(f_\theta(x_i)_k)}\right)$$

  where:
  - $f_\theta(x_i)_k$ is the predicted logit for class k,
  - y_i is the true class label for input x_i,
  - K is the total number of classes (in this case, 2).

This loss function is computed for each data point and averaged over the mini-batch during training.

2. Empirical Risk L:

The **empirical risk** L represents the average loss over the entire training set. It is the quantity the model is trying to minimize during training. In the case of mini-batch training (as in this code), the empirical risk is computed as the average cross-entropy loss over the data points in the mini-batch.

Mathematically, if there are n training examples, the empirical risk $L(\theta)$ is:

$$L(\theta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\theta(x_i), y_i)$$

where $l(f_\theta(x_i), y_i)$ is the cross-entropy loss for each example (x_i, y_i).

In the code, the empirical risk is minimized using stochastic gradient descent (SGD) as the optimizer:

```
optimizer = torch.optim.SGD(model.parameters(), lr=0.5)
```

This step updates the model parameters θ to minimize the empirical risk.

3. Probability Distribution p_θ (Log-Likelihood View):

In the **log-likelihood view**, the cross-entropy loss can be interpreted as minimizing the negative log-likelihood of the true class under a predicted probability distribution.

The logits output by the model, $f_\theta(x_i)$ are converted into class probabilities using the **softmax function**:

$$p_\theta(y_i|x_i) = \frac{\exp(f_\theta(x_i)_{y_i})}{\sum_{k=1}^{K} \exp(f_\theta(x_i)_k)}$$

where:

- $f_\theta(x_i)_k$ is the logit for class k.
- $p_\theta(y_i|x_i)$ is the predicted probability for the true class y_i.

The **log-likelihood** for a single data point is the log of the probability assigned to the correct class:

$$\log(p_\theta(y_i|x_i)) = \log\left(\frac{\exp(f_\theta(x_i)_{y_i})}{\sum_{k=1}^{K} \exp(f_\theta(x_i)_k)}\right)$$

The **negative log-likelihood** is what the cross-entropy loss computes, and by minimizing this loss, the model maximizes the likelihood of the correct class under the predicted probability distribution. The loss function and the empirical risk are thus directly related to the likelihood function of the true labels given the model's predictions.

In [10]:
```
import torch.nn.functional as F


torch.manual_seed(123)
model = NeuralNetwork(num_inputs=2, num_outputs=2)

device = torch.device("cuda" if torch.cuda.is_available() else "cpu") # NEW
model = model.to(device) # NEW

optimizer = torch.optim.SGD(model.parameters(), lr=0.5)

num_epochs = 3

for epoch in range(num_epochs):

    model.train()
    for batch_idx, (features, labels) in enumerate(train_loader):

        features, labels = features.to(device), labels.to(device) # NEW
        logits = model(features)
        loss = F.cross_entropy(logits, labels) # Loss function

        optimizer.zero_grad()
        loss.backward()
        optimizer.step()

        ### LOGGING
        print(f"Epoch: {epoch+1:03d}/{num_epochs:03d}"
              f" | Batch {batch_idx:03d}/{len(train_loader):03d}"
              f" | Train/Val Loss: {loss:.2f}")

    model.eval()
    # Optional model evaluation
```

```
Epoch: 001/003 | Batch 000/002 | Train/Val Loss: 0.75
Epoch: 001/003 | Batch 001/002 | Train/Val Loss: 0.65
Epoch: 002/003 | Batch 000/002 | Train/Val Loss: 0.44
Epoch: 002/003 | Batch 001/002 | Train/Val Loss: 0.13
Epoch: 003/003 | Batch 000/002 | Train/Val Loss: 0.03
Epoch: 003/003 | Batch 001/002 | Train/Val Loss: 0.00
```

In [11]:
```
def compute_accuracy(model, dataloader, device):

    model = model.eval()
    correct = 0.0
    total_examples = 0

    for idx, (features, labels) in enumerate(dataloader):

        features, labels = features.to(device), labels.to(device) # New

        with torch.no_grad():
            logits = model(features)

        predictions = torch.argmax(logits, dim=1)
        compare = labels == predictions
        correct += torch.sum(compare)
        total_examples += len(compare)

    return (correct / total_examples).item()
```

In [12]:
```
compute_accuracy(model, train_loader, device=device)
```

Out[12]:
```
1.0
```

In [13]:
```
compute_accuracy(model, test_loader, device=device)
```

Out[13]:
```
1.0
```

A.9.3 Training with multiple GPUs

See DDP-script.py