TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2024

Minibatching: The Batch Index

Minibatching

We run some number of instances together (or in parallel) and then do a parameter update based on the average gradients of the instances of the batch.

For NumPy minibatching is not so much about parallelism as about making the vector operations larger so that the vector operations dominate the slowness of Python. On a GPU minibatching allows parallelism over the batch elements.

Minibatching

With minibatching each input value and each computed value is actually a batch of values.

We add a batch index as an additional first tensor dimension for each input and computed node.

Parameters do not have a batch index.

The Batch Index

We let B be the number of elements in a single minibatch and let b, with $0 \leq b \leq B - 1$, be an index naming a particular element of the current minibatch.

MLP with a Batch Index

b— batch index, i — input feature index

j — hidden layer index, $\hat{y}$ — possible label

$$\Phi = (W^0[j,i],\ C^0[j],\ W^1[\hat{y},j],\ C^1[\hat{y}])$$

$$h[b,j] = \sigma\left(\sum_i \left(W^0[j,i]\ x[b,i]\right) - C^0[j]\right)$$

$$s[b,\hat{y}] = \sigma\left(\sum_j \left(W^1[\hat{y},j]\ h[b,j]\right) - C^1[\hat{y}]\right)$$

$$P_\Phi[b,\hat{y}] = \underset{\hat{y}}{\text{softmax}}\ s[b,\hat{y}]$$

Backpropagation with Minibatching

$$\text{for } b, i, j \quad \tilde{h}[b, j] \mathrel{+}= W[j, i]\, x[b, i]$$

$$\text{for } b, i, j \quad x.\text{grad}[b, i] \mathrel{+}= \tilde{h}.\text{grad}[b, j]\, W[j, i]$$

$$\text{for } b, i, j \quad W.\text{grad}[j, i] \mathrel{+}= \frac{1}{B}\, \tilde{h}.\text{grad}[b, j]\, x[b, i]$$

B is the number of batch elements. By convention parameter gradients are averaged over the batch.

Setting the Batch Size

The batch is typically made as large as the hardware will support.

Theoretically extremely large batches can make SGD require more epochs and hence more energy consumption even for parallel batch processing.

But empirically, unless one has thousands of GPUs, it seems the batch can be as large as possible without requiring additional epochs.

END