1. Fully connected Layer

* 此區塊的forward，就wx + b，這邊使用torch的matrix multiplication去將weight和input相乘，最後再加上bias。
* Text

  Description automatically generated在backward的部分，因為需要更新w和b所以要分別對這2個進行微分，並且乘以upstream的backprop。dw的部分是將wx + b對w微分因此就剩下x也就是input，將input乘以backprop得到dw，但因為是mini batch optimization，所以必須去取input size的平均。wx + b對b的微分是1，所以db就會等於input去除以batch size。dx 的部分是對x作微分，所以會剩下w，再乘以backward得到所求。

1. ReLu

* 在ReLu forward的部分，要將小於0的部分都轉換成0，大於0的部分則不變，因此透過torch.maximum來實現。
* Text

  Description automatically generated在backward的部分，如果在forward時input大於0就會輸出input本身，因此微分後就會等於1; 如果input 小於等於0的話，就會輸出0，所以微分後也會等於0。最後再將他與backprop做element wise multiplication

1. Softmax\_CrossEntropy

* Forward : 首先將每張照片的10個神經元輸出減去裡面的最大值，此步驟是為了增加numerical stability，防止接下來使用exponential產生特別大的值。接下來則是將每個數值取exponential，並除以10個exponential的sum來產生每個神經元的probability。接下來就是找到最大機率的神經元的index，並轉換成one hot的形式以利計算loss。接下來，就是計算cross entropy loss，將產生的one hot乘以log probability，最後再將batch中所有的cross entropy取平均並加上負號。
* Text

  Description automatically generatedBackward : 在forward時是做softmax及cross entropy loss，所以針對這些公式去進行微分會得到probability減去one hot型式的label，再除以batch size。

1. SGD

* Text

  Description automatically generated此optimizer是使用前面所計算參數的gradient來更新參數，在SGD中只需要將原本參數值減去gradient乘 一個學習率。

1. SGD + Momentum

* Text

  Description automatically generated此optimizer除了SGD以外還加了momentum，透過momentum可以降低每個mini batch的noise，較不會因為單一batch而偏離主要行駛的方向。主要是使用移動平均的方式去計算，由圖中可以看到我們透過velocity去記錄過去的更新方向，並將期程以一個momentum使得越接近的update佔較大的比重。

1. Adam ( Adative momentum estimation )

* 將momentum與RMSProp結合起來。 ; ;
* AdaGrad ( Adaptive Gradient ) 可以自動調整learning rate，其更新方式為，透過此方式每個weight都會有不同的learning rate。在真實世界，有些feature是sparse(大部分的feature是0)、有些feature是dense(大部分的features是非0)，在AdaGrad中可以針對sparse的部份給他較高的learning rate，dense的部分則給他較低的learning rate。
* RMSProp ( Root Mean Squared Propagation ) : ; ，RMSProp有點類似AdaGrad，但是他使用moving average，所以可以避免在多個iteration下step size過小。
* Bias correction的部分主要是因為momentum和velocity的初始值都是0，所以會有偏向0的傾向，因此會將他們除以，隨著t變大， 就會接近1

Text

Description automatically generated

1. Test Optimizer

* Chart, histogram

  Description automatically generated在這邊可以看到依照設定的參數來比較可以發現Adam可以快速降低loss，並且在成效上也比其他的Optimizer好，因此選擇Adam

Chart

Description automatically generated

1. Training with Best Hyperparameters

* Chart, line chart, histogram

  Description automatically generatedGraphical user interface, text

  Description automatically generated這邊跑出的結果為43.63%

Text

Description automatically generated