**Inline Question 1:**

描述一下这个实验的结果。权重初始化的规模如何影响 带有/没有Batch Norm的模型，为什么?

**Answer:**

使用Batch Norm时，即使我们使用了不好的权重初始化的规模，我们仍然可以得到一个不错的结果；但是如果没有Batch Norm，我们对权重初始化尺度非常敏感，即使用了不好的权重初始化规模会使得结果变得难以接受。

使用Batch Norm可以尽可能地使分布保持在原有的状态,所以使用Batch Norm最终可以让模型对权重初始化的规模不那么敏感。

当权重初始化较大时，由于baseline 依然逐个normalization，导致训练更不可控，而batchnorm 依然能保持和小权重时差不多的loss 水平，因为batch 后不易受受个体影响。同时可以看出，batchnorm整体优于basenorm，这是因为深层神经网络内部特征的转移分布可能会使训练深层网络更加困难，而batchnorm 则解决了这一点。

## Inline Question 2:

描述一下这个实验的结果。请问Batch Norm和batch size之间的又什么关系？为什么会出现这种关系？

## Answer:

使用Batch Norm时，在训练集上，batch size越大，模型的效果越好；在验证集上虽然效果较为接近，不过batch size越大，模型的效果还是会有一些提高。因为Batch Norm可以使分布保持在原有的状态，从而降低随机性

batch size 越大时，batch norm 的模型准确率更高，因为这样会有更快的收敛速度和并行能力。只要硬件资源足够，并且模型没明显过拟合，尽量把batch size 设大。batch size越小，batch normalization越差。当使用非常小的batch size时，baseline模型可能优于batch normalization模型。 出现这个问题是因为当我们计算一个批次的均值和方差时，试图找到整个数据集的均值以及方差的近似值。 因此，对于小批量，这些值可能会非常嘈杂。 而对于大批量，我们可以获得更好的均值和方差的近似。

**Inline Question 3:**

下面的数据预处理步骤中，哪些类似于Batch Norm，哪些类似于Layer Norm?

1. Scaling each image in the dataset, so that the RGB channels for each row of pixels within an image sums up to 1.
2. Scaling each image in the dataset, so that the RGB channels for all pixels within an image sums up to 1.
3. Subtracting the mean image of the dataset from each image in the dataset.
4. Setting all RGB values to either 0 or 1 depending on a given threshold.

**Answer:**

Batch norm 是按样本批量，Layer Norm 按通道/时间片(RNN)/隐层节点数量(MLP) 批量按特征 数做norm.

1.Batch norm，对样本批量

2.Layer Norm，因为是通道数

3.Batch norm，对样本进行的均值化

4.Layerorm，对通道值均值化。

**Inline Question 4:**

什么时候layer normalization可能不工作（不起作用），为什么?

1. 在非常深的网络上使用
2. 特征的维度非常的小
3. 有非常高的正则化项

**Answer:**

2 特征的维度非常的小会使得layer normalization可能不工作（不起作用）。

因为layer normalization是用图片的特征向量的所有元素算出方差、均值，如果特征的维度非常小的话就会导致方差、均值误差较大。

3 有非常高的正则化项也可能使得layer normalization不起作用

因为，非常高的正则化项意味着惩罚权重的程度很大，输出值的大小会减小，layer normalization的作用被削减。