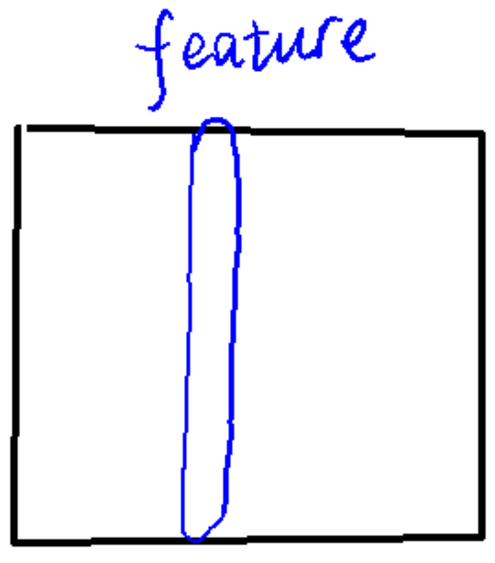
Batch-Norm



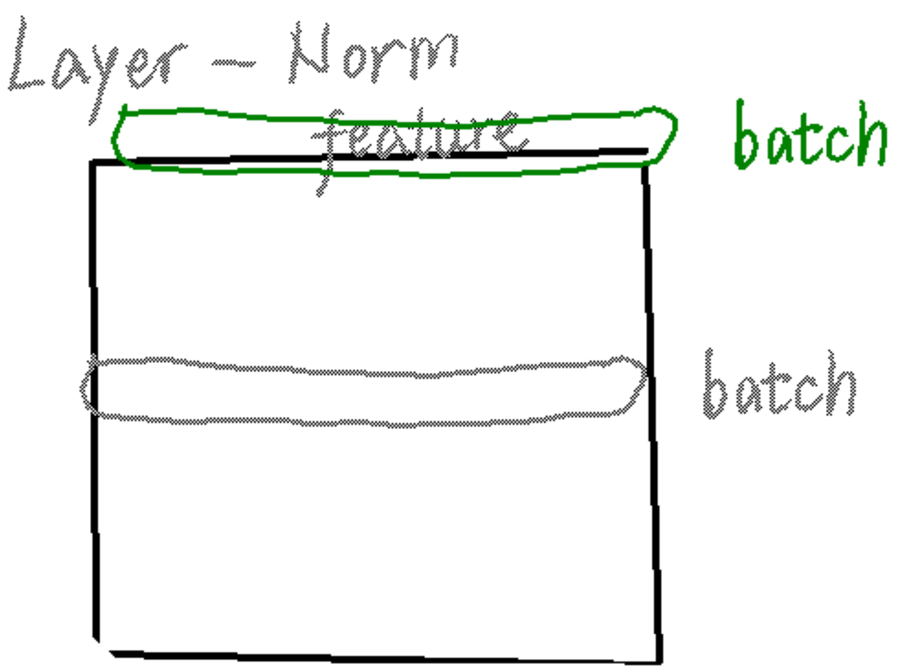
每次把每个列(每个特征)
在一个mini-batch里，均值变成0，方差变为1
(训练)

把全局均值和方差计算出（预测时）

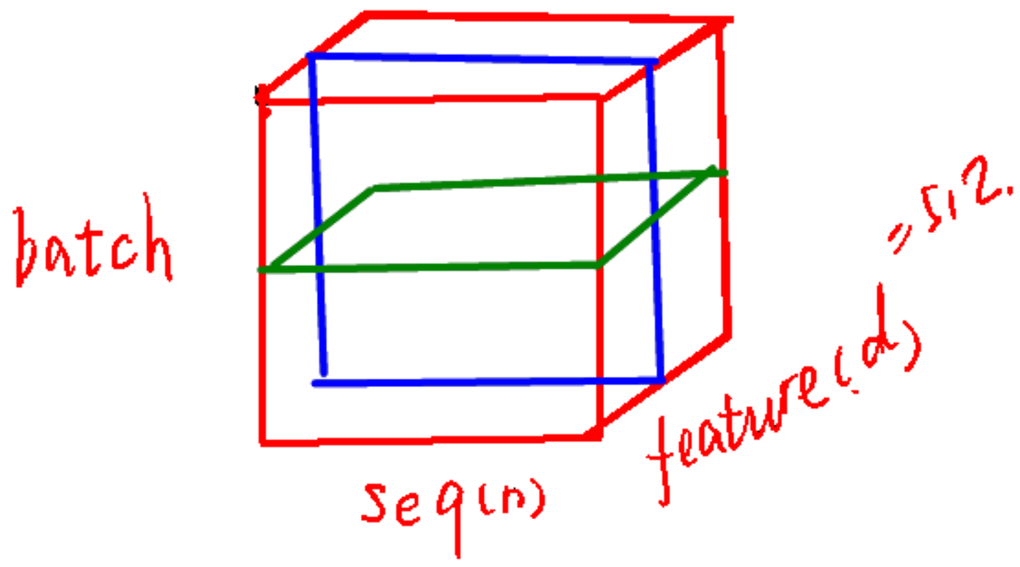
同时学习 λ 和 β，通过学习可把向量放成任意均值，方差的东西.

Layer - Norm
feature

把每行(每个样本)作 Normalization.



把每行(每个样本)作 Normalization.

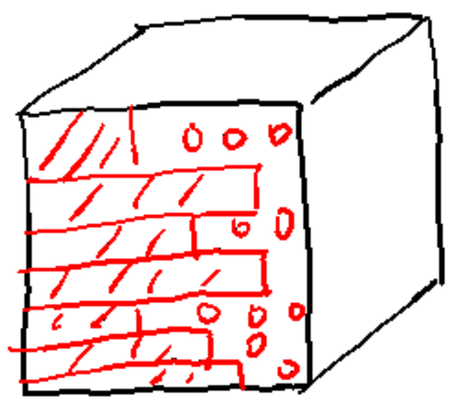


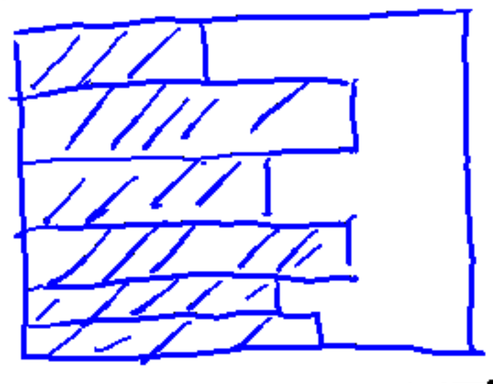
每个词对应一个 seq.

蓝色：BatchNorm

绿色：LayerNorm

实际情况，seq长度可能会变化



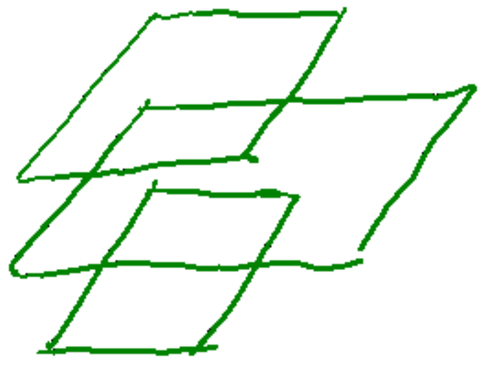


← Batch-Norm

样本长度变化大时

Mean, Var 抖动大

→按见过 同时，预测时用全局的 Mean 和 Var



← Layer Norm

每个样本计算 Mean 和 Var

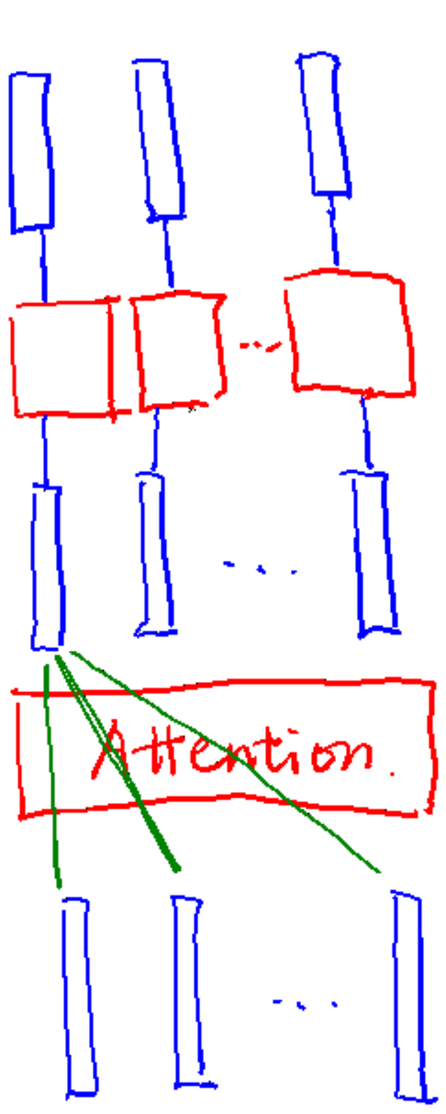
也不需要全局的 Mean 和 Var

Same: RNN 和 Transformer 都是用一个 MLP / 线性层做语义空间转换

Difference: 如何传递序列信息

RNN = 上时刻 → 下时刻

Transformer: Attention层，全局信息 → MLP转换



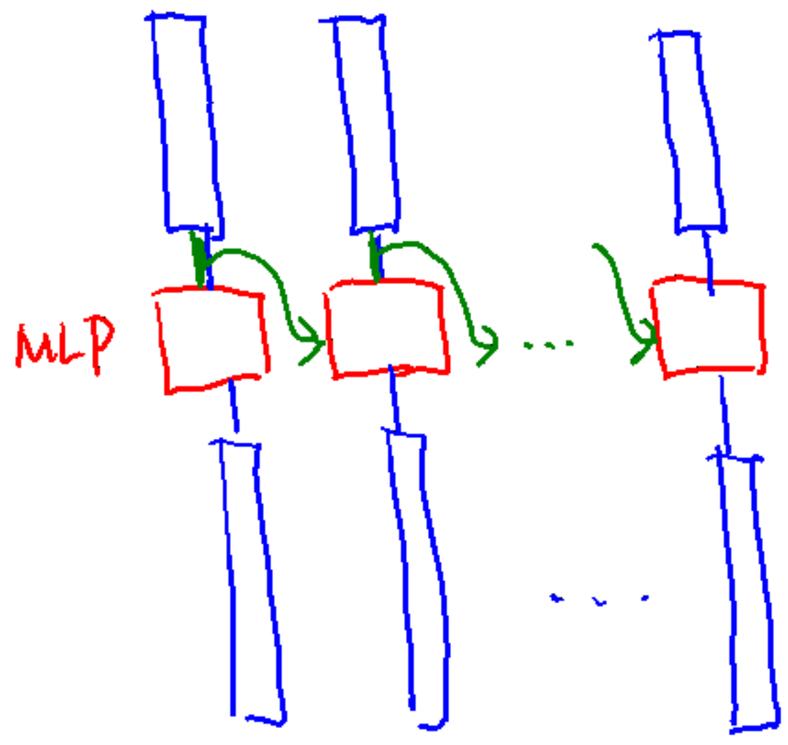
权重一样

MLP - 可分开做

↑ ∴ MLP可分开.

加权 序列信息汇聚完成

Transformer



RNN