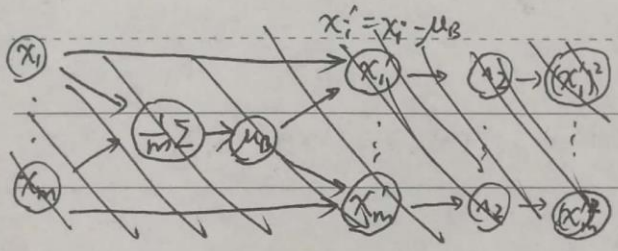
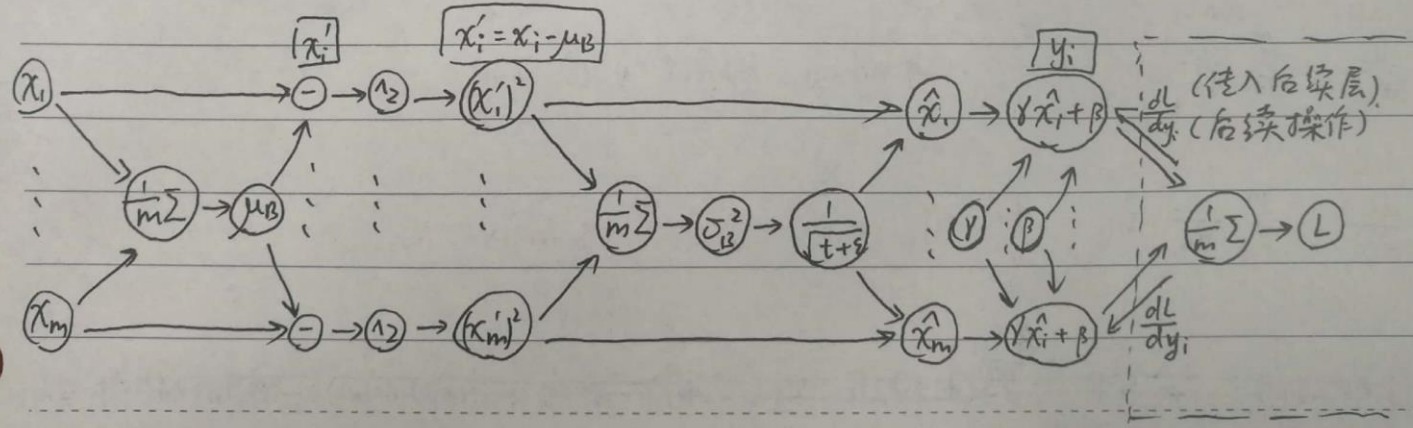
① Normalization

- Batch normalization、Layer normalization、Spatial normalization、Group normalization.
区别仅在于用在标准化上的维度不同，写出BN的BP后，对矩阵维度进行转置直接调BN包即可实现后3者（详见代码）

先画出m个样本的（一维）computation graph：





从 $\frac{dL}{dy_i}$ 开始，进行BP，可参照原论文ppt对求导正确性。（红笔向量化）

$\frac{dL}{d\hat{x}_i}$ (N×D) = $\frac{dL}{dy_i}$ (N×D) · γ (1×D) ⇒ broadcast 即可

$\frac{dL}{d\sigma_B^2}$ (1×D) $= \sum_i \frac{dL}{d\hat{x}_i} \cdot \frac{d\hat{x}_i}{d\sigma_B^2} = \sum_{i=1}^{m} \frac{dL}{d\hat{x}_i} \left(-\frac{1}{2} \frac{1}{\sigma_B^2+\varepsilon} \cdot \hat{x}_i\right)$ ⇒ 括号内 broadcast，$\sum_{i=1}^{m}$ · 用 $(N\times D)^T(N\times D)$ 实现，最后取对角

$\frac{dL}{dx_i'}$ (N×D) $= \frac{dL}{d\hat{x}_i} \cdot \frac{d\hat{x}_i}{dx_i'} + \frac{dL}{d\sigma_B^2} \cdot \frac{d\sigma_B^2}{dx_i'} = \frac{dL}{d\hat{x}_i} (N\times D) \cdot \frac{1}{\sqrt{\sigma_B^2+\varepsilon}} (1\times D) + \frac{dL}{d\sigma_B^2} (1\times D) \cdot \frac{1}{m} 2x_i' (N\times D)$ ⇒ broadcast 即可。

$\frac{dL}{d\mu_B}$ (1×D) $= \sum_i \frac{dL}{dx_i'} \cdot \frac{dx_i'}{d\mu_B} = \sum_{i=1}^{m} \frac{dL}{dx_i'} (N\times D) \cdot (-1)$ ⇒ $\sum_{i=1}^{m}$ · 用 np.ones(N).dot(N×D) 实现。

$\frac{dL}{dx_i}$ (N×D) $= \frac{dL}{d\mu_B} \cdot \frac{d\mu_B}{dx_i} + \frac{dL}{dx_i'} \cdot \frac{dx_i'}{dx_i} = \frac{1}{m} \frac{dL}{d\mu_B} (1\times D) + \frac{dL}{dx_i'} (N\times D)$ ⇒ broadcast 即可。

$\frac{dL}{d\beta}$ (1×D) $= \sum_{i=1}^{m} \frac{dL}{dy_i}$ (N×D) ⇒ $\sum_{i=1}^{m}$ · 用 np.ones(N).dot(N×D) 实现

$\frac{dL}{d\gamma}$ (1×D) $= \sum_{i=1}^{m} \frac{dL}{dy_i}$ (N×D) · $\hat{x}_i$ (N×D) ⇒ $\sum_{i=1}^{m}$ · 用 $(N\times D)^T(N\times D)$ 实现，最后取对角

~~关于取对角：σ_B^2 与 γ 为特征独有，相互之间不影响，$(N\times D)^T(N\times 1)$ 得到的非对角项代表~~ 各个

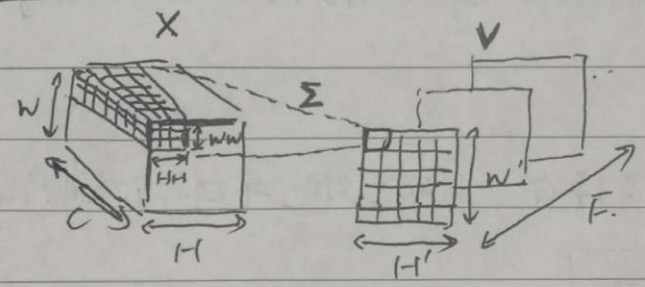
关于取对角：等价于两矩阵作 element-wise 的乘法，再用向量化实现 $\sum_{i=1}^{m}$ ·

即 $\frac{dL}{d\gamma} = np.ones(N1).dot\left(\frac{dL}{dy} * \hat{x}\right)$

$\left(= np.ones(N1).dot\left(\frac{dL}{dy} \odot \hat{x}\right).\right)$

$\frac{dL}{d\sigma_B^2}$ 同理

② CNN



前向传导较简单，反向传导需理清各变量依赖关系。

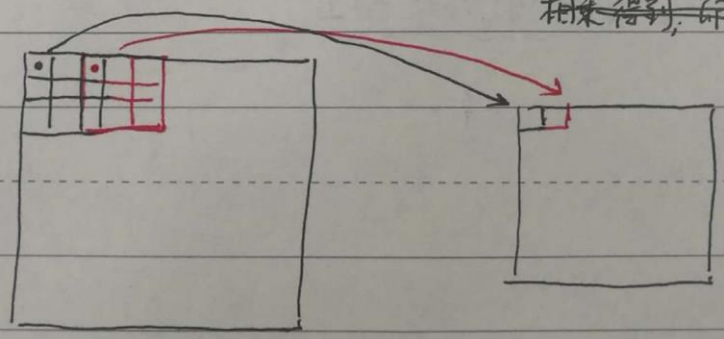
• 参数：卷积核 $W_F \in R^{HH \times WW \times C}$，可见其与F有关，所以先讨论 ∇F.

但 $W_F \in R^{HH \times WW \times C}$，注意到 C 维上行为是相同的，所以再简化，讨论 ∇C.

故 ∇F，∇C，$\frac{dL}{dw_{hw}} = \sum_{k,l} \frac{dL}{dv_{kl}} \cdot \frac{dv_{kl}}{dw_{hw}}$ （下标与w重了... 不想用orz）

$$= \sum_{l} \sum_{k} \left[\frac{dL}{dv_{kl}}\right] \cdot \left[\frac{dv_{kl}}{dw_{hw}}\right]$$

↙ 上层传入　　↘ 找出 V 中与 W_{hw} 有关的元素，~~而 V 中 ww 与 X 相乘得到，即寻找 X 中与 W_{hw}~~



由于 $V = X * W$，即在 X 中寻找与 W_{hw} 相乘过的元素

X　　V

因此 $\frac{dL}{dw} = X_s * \frac{dL}{dV}$ （* 为卷积操作）

其中 X_s 是按步长间隔取出的矩阵，但卷积步长为1

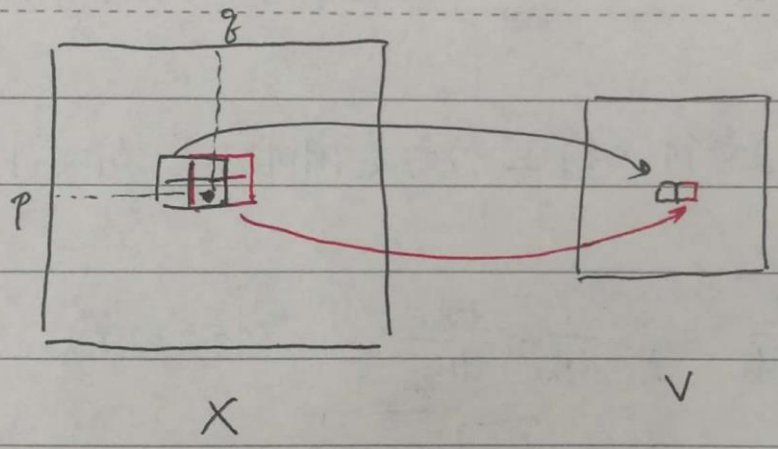
(2) $\frac{dL}{db} = \sum_{k,l} \frac{dL}{dV_{kl}} \cdot \frac{dV_{kl}}{db}$ ，直接将 $\frac{dL}{dV}$ 所有元素求和即可。

(3) 求 $\frac{dL}{dX}$，由于涉及多样本，此时先讨论单样本，再进行向量化

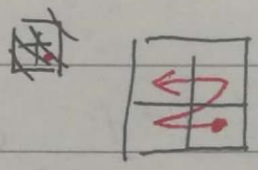
$$\frac{dL}{dx_{pq}^{(i)}} = \sum_{k,l} \frac{dL}{dV_{kl}} \cdot \frac{dV_{kl}}{dx_{pq}^{(i)}}$$

$\frac{dL}{dV_{kl}}$：上层传入

$\frac{dV_{kl}}{dx_{pq}^{(i)}}$：找出V中与 $x_{pq}^{(i)}$ 有关的元素



可见，与 $x_{pq}^{(i)}$ 相乘的W顺序是逆向的，即



（"逆"是相对卷积得出的值在V中的顺序而言）

因此 $\frac{dL}{dX} = rot180(W) * \frac{dL}{dV}$

其中卷积步长为1

(4) 边界情况

- 对 $\frac{dL}{dW}$ 无 0 填充. (但注意 $\frac{dL}{dW} = X_s * \frac{dL}{dV}$ 中, X_s 从填充后的 X 中取出, 即前向传导时用的 X).

- 对 $\frac{dL}{dX}$ 需在 $\frac{dL}{dV}$ 边缘及内部 0 填充.
  边缘填充 "W 的长度 -1" (或"宽度 -1"), 内部填充 "步长 -1"
  同时注意, 求出的 $\frac{dL}{dX}$ 是带填充的, 须去掉边缘

填充细节是通过分析 $X * W = V$ 第一行的行为总结出的. 很有意思, 留作思考. 方法类似 (1)、(3) 的画图分析.

- 这里给出 $\frac{dL}{dV}$ 填充后大小与 $rot180(W) * \frac{dL}{dV}$ 的大小

  设 $X \in R^{H \times W}$, $W \in R^{HH \times WW}$, $\frac{dL}{dV} \in R^{H' \times W'}$

  步长为 S, 前导时对 X 填充系数为 P.

  则填充后 X_pad 长为: ~~$(H+2P-HH)/S+1$~~ $H+2P$.

  按上述规则填充后 $\frac{dL}{dV}$_pad 长为: $(H'-1)(S-1)+H'+2(HH-1)$

  $rot180(W) * \frac{dL}{dV}$ 后 (注意步长始终为 1) 长为:

  $$(H'-1)S+1+2(HH-1)-HH+1 = (H'-1)S+HH$$

  ~~注意到 $W * X_{pad}$ 的长为 $H+2P$~~

  注意到 $W * X\_pad$ 的长为: $(H+2P-HH)/S+1 = H'$, 故

  $$H+2P = (H'-1)S+HH.$$

  可见 $\frac{dL}{dV}$ 填充后再卷积, 大小与 X_pad 相同, 这也说明反向传播前述的维度上是正确/合理的.