量化误差分析

Sensetime HPC Group

定义

- 为了研究方便，我们假定： $x_i \sim N(0,1)$ 且独立同分布

  - 我们定义量化函数 $Q(x,s) = \begin{cases} [\frac{x}{s}], & else \\ 127, & \frac{x}{s} > 127.5 \\ -127, & \frac{x}{s} < -127.5 \end{cases}$ ，额外定义截断值 $c = s * 127.5$
  - 我们定义反量化函数 $DQ(x,s) = x * s$
  - 我们定义x的量化值为 $x'(s) \triangleq DQ(Q(x,s),s)$
  - $L_{MSE} = \frac{\sum_i (x_i - x_i')^2}{N}$
  - $L_{N/S} = \frac{\sum_i \frac{(x_i - x_i')^2}{x_i^2}}{N}$

量化误差分析 - Bernard Widrow公式

$$E\{L_{MSE}\} = (x-c)^2P_1(x) - 2((x-c)P_2(x) - P_3(x))\Big|_{c+\frac{s}{2}}^{inf} + (x+c)^2P_1(x) - 2((x+c)P_2(x) - P_3(x))\Big|_{-inf}^{-c-\frac{s}{2}}$$

$$+(P_1(c+\frac{s}{2}) - P_1(-c-\frac{s}{2}))\frac{s^2}{12}$$

- 该等式是量化核心成果之一，该等式表明了量化误差是正负截断误差与表示误差的累计和。
- 对于高斯分布而言，截断误差随着截断值增长而指数级收敛；表示误差随着截断值增长而增长，增长速度为二次方级。
- 在截断值不变的情况下，表示误差随scale的增长而增长，增长速度为二次方级。

向前延伸一步：网络中的误差

考虑一个单层无激活的网络

$$Y = W_1 X + b_1$$

其中 Y 是网络输出结果，不失一般性我们说 $Y \in R^{K*K}$

$W_i \in R^{K*K}$ 是网络权重 $b_i \in R^K$ 为网络偏置项

量化误差的表示形式

考虑该网络的量化形式：

$$Y = W_1 X + b_1$$

引入量化函数$f(x, s)$

$$Y' = f(W_1, s_{W_1}) f(X, s_X) + b_1$$

我们引入量化函数 $f(x,s) = \begin{cases} [\frac{x}{s}]s, & else \\ 127s, & \frac{x}{s} > 127.5 \\ -127s, & \frac{x}{s} < -127.5 \end{cases}$

量化误差的表示形式

考虑该网络的量化形式：

$$Y = W_1 X + b_1$$

引入量化函数$f(x, s)$

$$Y' = f(W_1, s_{W_1}) f(X, s_X) + b_1$$

引入量化误差项E_{W_1}, E_X，令 $f(W_1, s_{W_1}) - E_{W_1} = W_1$

$$Y' = (W_1 + E_{W_1})(X + E_X) + b_1$$

量化误差的分布

$$Y' = (W_1 + E_{W_1})(X + E_X) + b_1$$

有以下结论：

$\mathrm{E}_{\mathrm{W}_1}$ 的分布取决于截断值的选取，取最大最小值截断时，服从$[\frac{-s}{2}, \frac{s}{2}]$上的均匀分布

取$\mathrm{k} - \sigma$截断时，绝大部分误差服从$[\frac{-s}{2}, \frac{s}{2}]$上的均匀分布，剩下的部分服从高斯分布。

问题：E_Y 是怎样的分布，与 $\mathrm{E}_{\mathrm{W}_1}$，$\mathrm{E}_{\mathrm{X}}$ 之间存在怎样的关系

输出误差从何而来？

$$Y' = (W_1 + E_{W_1})(X + E_X) + b_1$$

问题：E_Y 是怎样的分布，与 E_{W_1}，E_X 之间存在怎样的关系

$$E_Y = (Y' - Y)$$

输出误差从何而来？

$$Y' = (W_1 + E_{W_1})(X + E_X) + b_1$$

问题：E_Y 是怎样的分布，与 E_{W_1}，E_X 之间存在怎样的关系

$$E_Y = (Y' - Y)$$

$$Y' = (W_1 + E_{W_1})(X + E_X) + b_1$$

$$Y' = W_1X + E_{W_1}X + E_XW_1 + E_XE_{W_1} + b_1$$

$$E_Y = (E_{W_1}X + E_XE_{W_1}) + E_XW_1$$

$$E_Y = E_{W_1}X' + E_XW_1$$

量化误差的正态性

$$E_Y = E_{W_1} X' + E_X W_1$$

量化误差 = 权重误差 * 量化输入 + 输入误差 * 权重

$$E_{W_1} X' = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \begin{bmatrix} X'_{11} & X'_{12} \\ X'_{21} & X'_{22} \end{bmatrix}$$

$$(E_{W_1} X')_{11} = \sum_i E_{1i} * X'_{i1}$$

量化误差的正态性

$$(E_{W_1}X')_{11} = \sum_i E_{1i} * X'_{i1}$$

当E_{1i}, X'_{i1}独立，且E_{W_1}与X'中的元素独立同分布，则根据中心极限定理可得：

$$(E_{W_1}X') \sim N(\mu_{E_{W_1}X'}, n\sigma^2_{E_{W_1}}\sigma^2_{X'})$$

注意适用条件：E_{W_1}
X'中的元素独立同分布，且求和量足够多

中心极限定理不适用的情况

1. 网络宽度过低

depthwise 中参与求和的量太少，不适用中心极限定理研究量化噪声的分布。

量化噪声的分布不会稳定收敛于高斯分布。

2. 数值分布不满足假设

RepVGG 类的网络存在将多个卷积的参数合并到一个卷积内的操作。

这导致 E_{W_1}, X' 等矩阵的元素不满足独立同分布的假设。

零均值量化误差

$$E_Y = E_{W_1}X' + E_X W_1$$

$$(E_{W_1}X') \sim N(\mu_{E_{W_1}X'}, n\sigma^2_{E_{W_1}}\sigma^2_{X'})$$

$$(E_X W_1) \sim N(\mu_{E_X W_1}, n\sigma^2_{E_X}\sigma^2_{W_1})$$

为了取得良好的量化效果，引入一个直流项 bias，使得 E_Y均值为0

$$E_Y = E_{W_1}X' + E_X W_1 + bias$$

这一方法也叫作 Bias Correction

误差方差分解

$$E_Y = E_{W_1}X' + E_X W_1$$

$$D\{E_Y\} = D\{E_{W_1}X' + E_X W_1\}$$

$$D\{E_Y\} = D\{E_{W_1}X'\} + D\{E_X W_1\} + COV(E_{W_1}X', E_X W_1)$$

信号误差期望

$$D\{E_Y\} = D\{E_{W_1}X'\} + D\{E_X W_1\} + COV(E_{W_1}X', E_X W_1)$$

$$\sigma_Y^2 = n\sigma_{E_X}^2\sigma_{W_1}^2 + n\sigma_{E_{W1}}^2\sigma_{X'}^2 + COV(E_{W_1}X', E_X W_1)$$

$$E\{SNR\} = \sum_i \frac{(Y_i - Y_i')^2}{NY_i{}^2} = \frac{\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{E_Y}}}\int_{-inf}^{inf} u^2\, e^{\frac{u^2}{2\sigma_{\mathrm{E_Y}}^2}} du}{\frac{1}{\sqrt{2\pi}\sigma_{\mathrm{Y}}}\int_{-inf}^{inf} v^2\, e^{\frac{(v-\mu_v)^2}{2\sigma_{\mathrm{Y}}^2}} dv} = \frac{\sigma_{\mathrm{E_Y}}^2}{\sigma_{\mathrm{Y}}^2 + \mu_{\mathrm{Y}}^2}$$

引入激活函数

$$Y' = (W_1+E_{W_1})(X + E_X) + b_1$$

是一个单层网络，考虑将结论扩展到多层，补入激活函数，得

$$Y_n' = u((W_1+E_{W_n})(Y_{n-1} + E_{Y_{n-1}}) + b_n)$$

$$\sigma^2_{E_{Y_n}} = v(n\sigma^2_{E_{Y_{n-1}}}\sigma^2_{W_n} + n\sigma^2_{E_{W_n}}\sigma^2_{Y_{n-1}'} + COV(E_{W_n}Y_{n-1}', E_{Y_{n-1}}W_n))$$

误差传播公式 - 1

$$\sigma^2_{E_{Y_n}} = v(\underbrace{n\sigma^2_{E_{Y_{n-1}}}\sigma^2_{W_n}}_{\text{Accumulated Error}} + \overbrace{n\sigma^2_{E_{W_n}}\sigma^2_{Y_{n-1}'}}^{\text{Immediate Error}} + \underbrace{COV(E_{W_n}Y_{n-1}', E_{Y_{n-1}}W_n)}_{\text{Cross Term}})$$

$$\sigma^2_{E_{Y_n}} = v(n\sigma^2_{E_{Y_{n-1}}}\sigma^2_{W_n} + n\sigma^2_{E_{W_n}}\sigma^2_{Y_{n-1}'} + COV(E_{W_n}Y_{n-1}', E_{Y_{n-1}}W_n))$$

由网络结构决定

网络中的误差传播

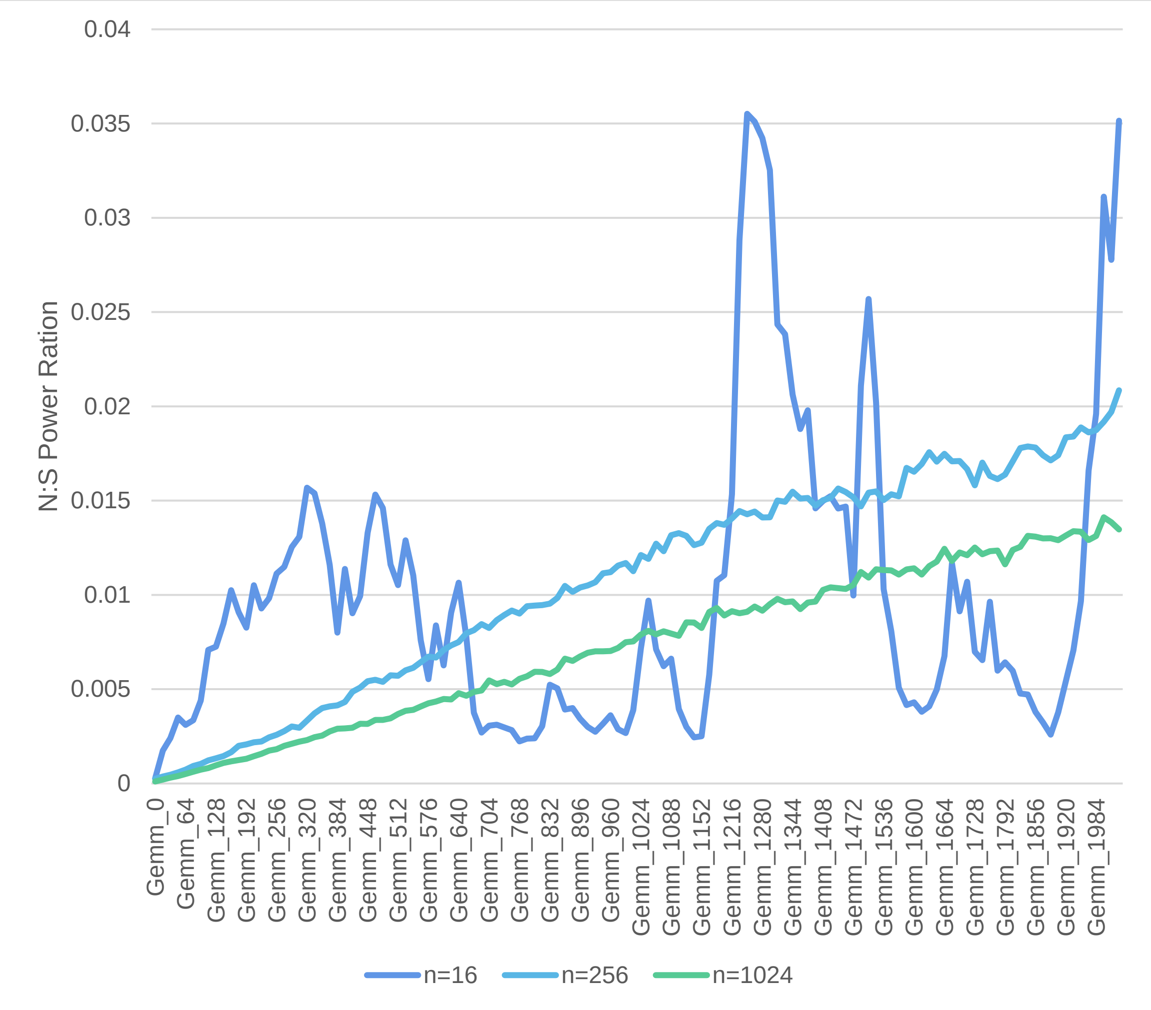
定义一连续网络结构:

$$Y_n{}' = u((W_1 + E_{W_n})(Y_{n-1} + E_{Y_{n-1}}) + b_n)$$

共计128层，定义参数k为每一层的神经元数量（网络宽度），定义u为激活函数。

额外地，我们将每一层的 Y_n 归一化，即进行 $Y_n = \frac{Y_n - \mu_{Y_n}}{\sigma_{Y_n}}$

网络中的误差传播



左图讨论了 n 对误差传播的影响，可以得到如下结论：

1. 在没有激活函数时，网络误差传播是线性的，

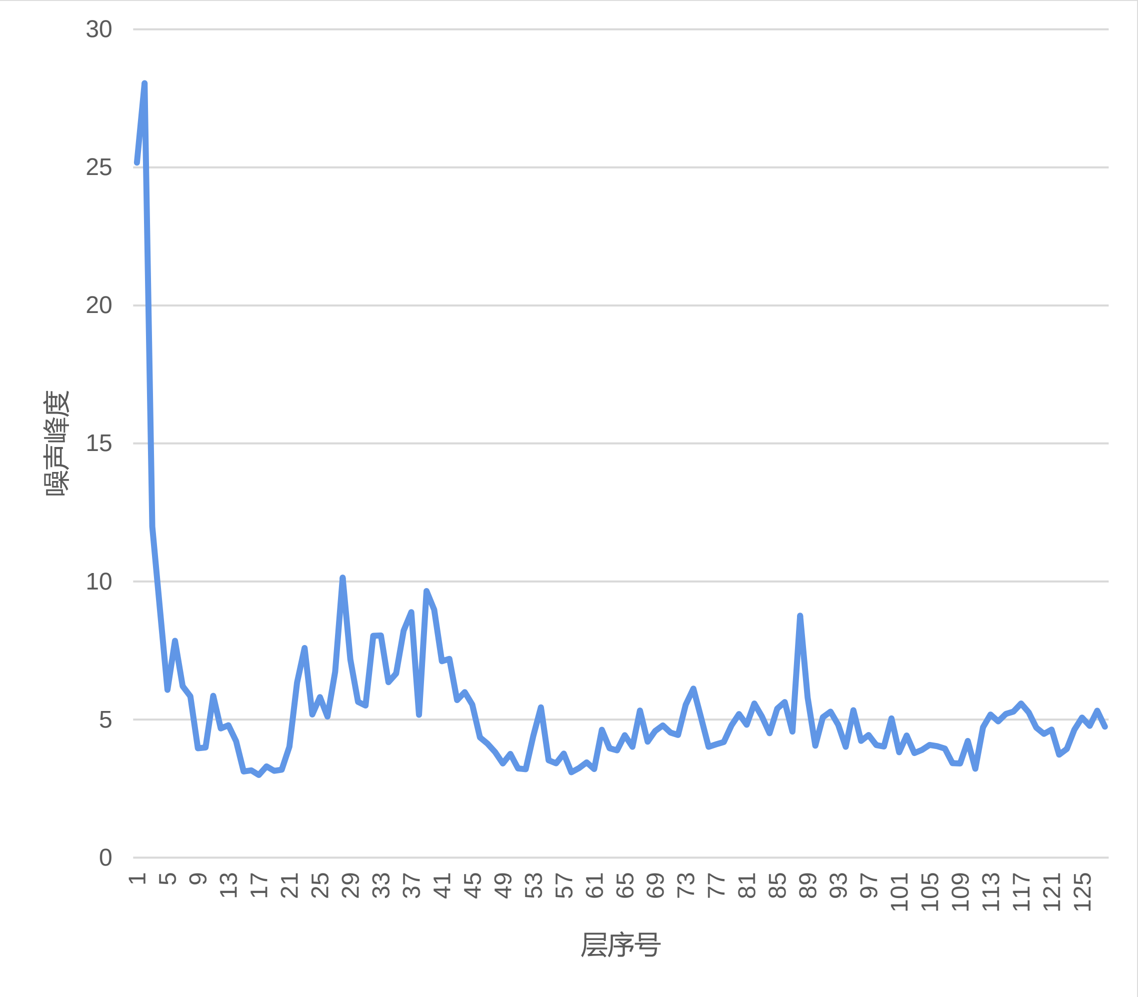
$$n\sigma_{W_n}^2 = 1$$

也就是说量化误差随着层数增加而线性增长

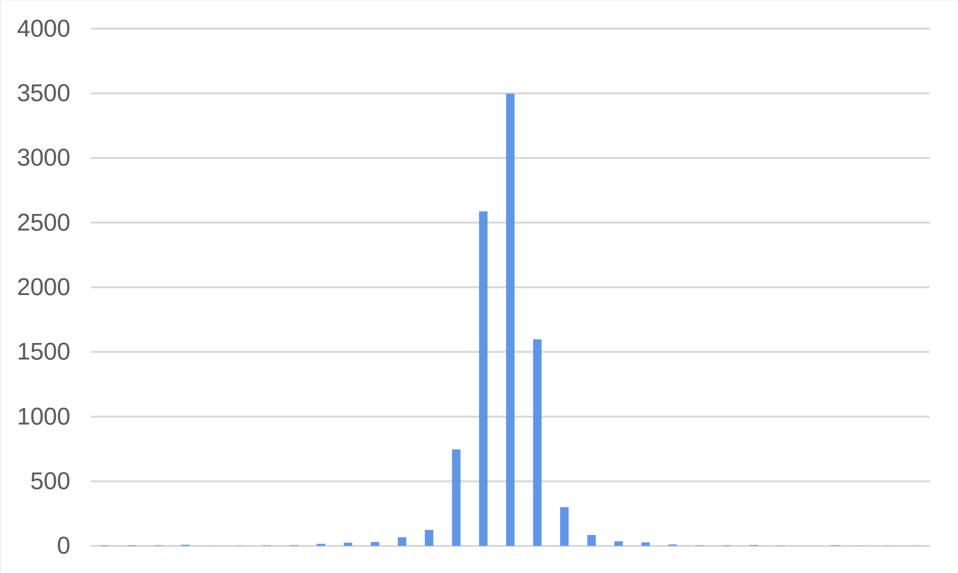
2. n过小则误差不稳定

3. 网络越宽，则局部量化误差越小

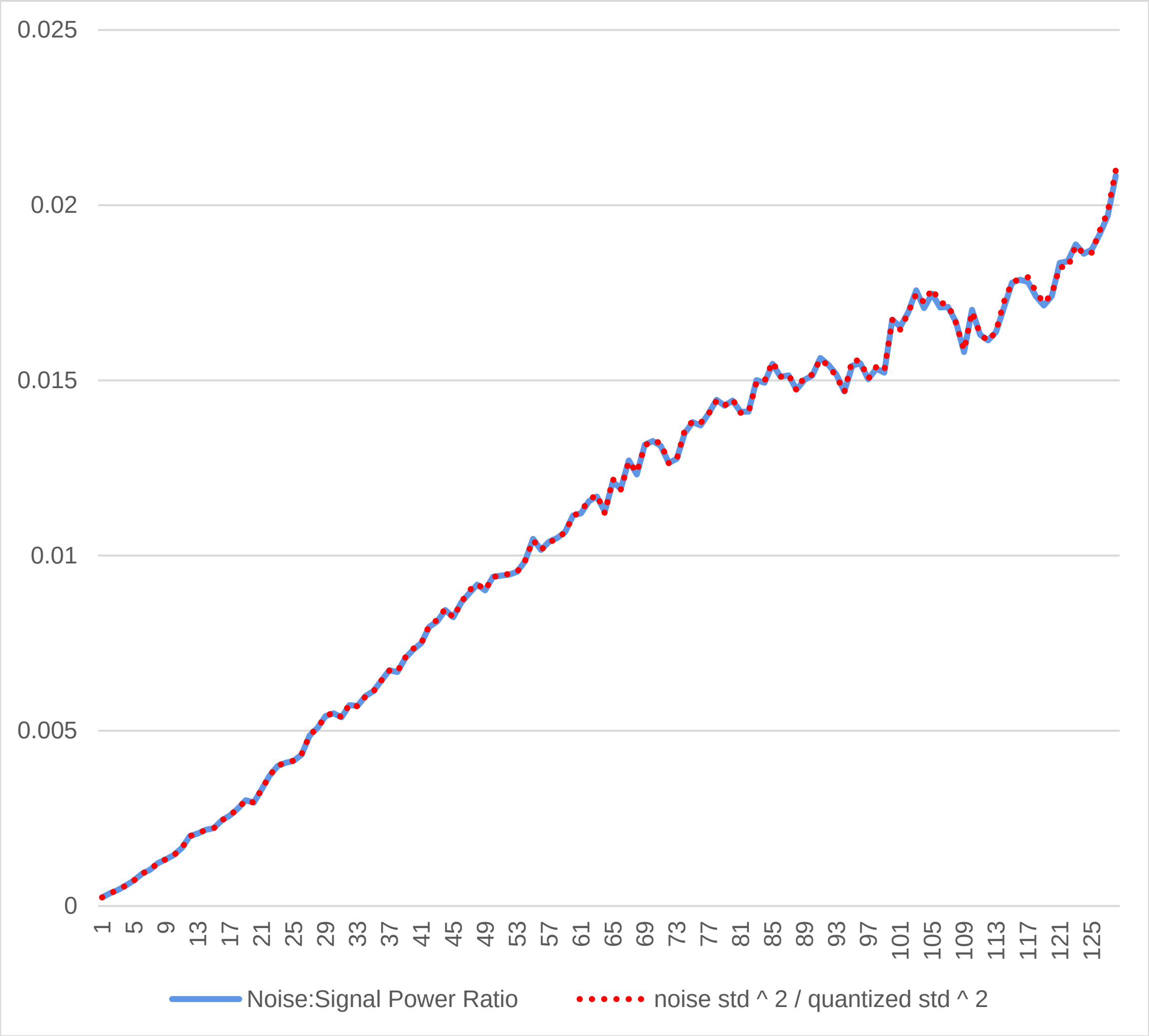
网络中的误差传播



随着层数增长，量化噪声趋于统计意义上的高斯分布。左图展示了随着层数增长，噪声分布的峰度变化情况。下图则展示了第一层时的噪声分布直方图：



网络中的误差传播



进一步地，我们考察

$$E\{SNR\} = \frac{\sigma_{\mathrm{E_Y}}^2}{\sigma_{\mathrm{Y}}^2 + \mu_{\mathrm{Y}}^2}$$

实际的 SNR 情况与分析情况完全一致

网络中的误差传播

总结：

1. 在没有激活函数时，且权重分布均匀时（$n\sigma_{W_n}^2 = 1$），网络误差传播是线性的， 量化误差随着层数增加而线性增长。

2. 网络越宽，则局部量化误差越小

3. 量化噪声随层数增长而趋近于高斯分布

4. $E\{SNR\} = \frac{\sigma_{\mathrm{E_Y}}^2}{\sigma_{\mathrm{Y}}^2 + \mu_{\mathrm{Y}}^2}$

思考为什么$n\sigma_{W_n}^2 = 1$

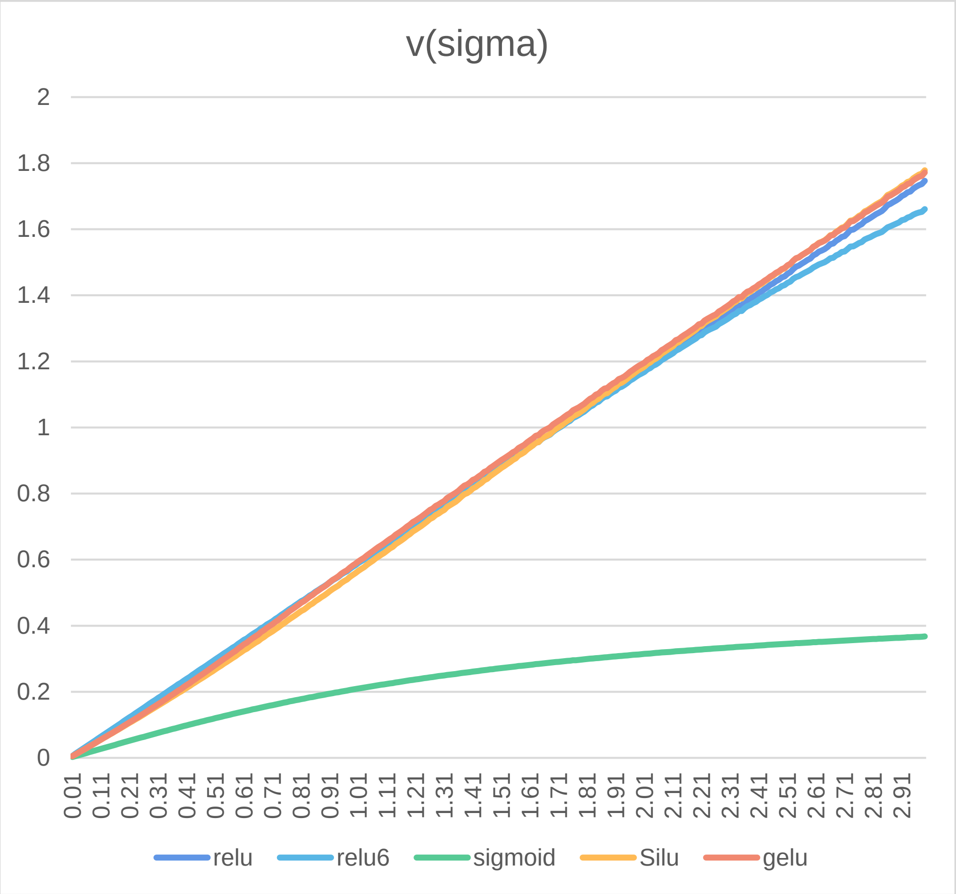
激活函数的影响

$$Y_n' = u(x \sim N(0, \sigma))$$

$$\sigma^2_{E_{Y_n}} = v(\sigma)$$

$$v(\sigma) = \int_{-inf}^{inf} (u(x) - \mu')^2 p(x) dx = \int_{-inf}^{inf} (u(x) - \int_{-inf}^{inf} u(x) p(x) dx)^2 p(x) dx$$

最优量化激活函数



思考：

- 左图展示了$\sigma' = v(\sigma)$的函数关系，但我们要求的是：

$$E\{SNR\} = \frac{{\sigma'}^2_{E_Y}}{{\sigma'}^2_Y + \mu_Y^2}$$

- 什么样的激活函数能够使得噪声影响降低？

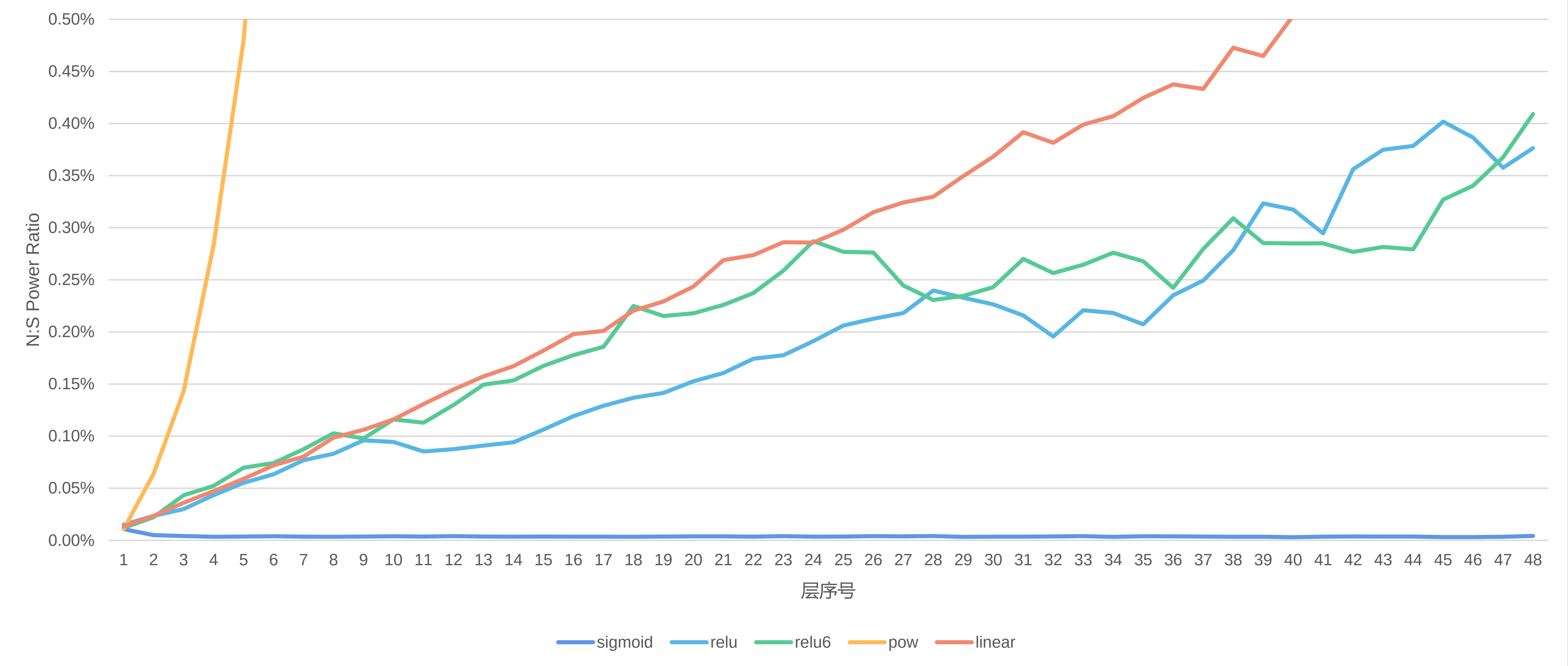
A. $y = clip([\frac{x*128}{6}] * \frac{6}{128}, 0, 6)$

B. $y = x^2$

C. $y = sigmoid(x)$

D. y=x

最优量化激活函数



0.50%
0.45%
0.40%
0.35%
0.30%
0.25%
0.20%
0.15%
0.10%
0.05%
0.00%
N:S Power Ratio
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48
层序号
sigmoid
relu
relu6
pow
linear