Invertibility of CNNs

Based on :

Gilbert, Anna C., Yi Zhang, Kibok Lee, Yuting Zhang, and Honglak Lee. "Towards Understanding the Invertibility of Convolutional Neural Networks." *arXiv preprint arXiv:1705.08664* (2017).

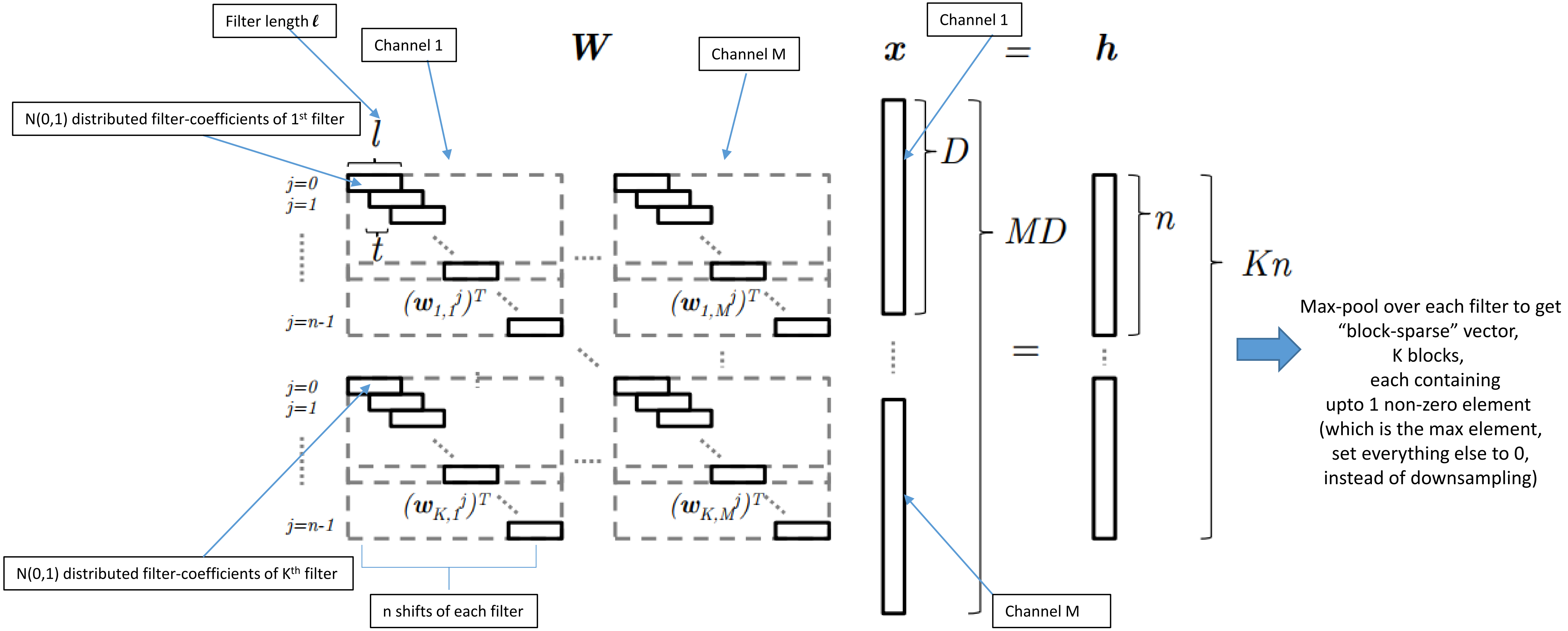

Figure 1: One-dimensional CNN architecture where $\boldsymbol{W} \in \mathbb{R}^{Kn \times MD}$ is the matrix instantiation of convolution over M channels with a filter bank consisting of K different filters. Note that a filter bank has K filters of size $l \times M$, such that there are lMK parameters in this architecture.

1-layer CNN

Step 1 : W x = h

Step 2 : $\hat{z} = maxpool(h, k)$ where $k \leq K$ (this is said to be model-sparse)

Hypothesis/reconstruction strategy:

We should be able to invert by using strategy:

$$\hat{x} = W^T \hat{z}$$

As long as W is constructed in *the "right way"*, we can show that $\hat{x} \cong x$

Theorem 3.3. *We assume that $\boldsymbol{W}^T$ satisfies the $\mathcal{M}_k^2$-RIP with constant $\delta_k \leq \delta_{2k} < 1$. If we use $\boldsymbol{W}$ in a single layer CNN both to compute the hidden units $\hat{\boldsymbol{z}}$ and to reconstruct the input $\boldsymbol{x}$ from these hidden units as $\hat{\boldsymbol{x}}$ so that $\hat{\boldsymbol{x}} = \boldsymbol{W}^T \mathbb{M}(\boldsymbol{W}\boldsymbol{x}, k)$, the error in our reconstruction is*

$$\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \leq \frac{5\delta_{2k}}{1 - \delta_k} \frac{\sqrt{1 + \delta_{2k}}}{\sqrt{1 - \delta_{2k}}} \|\boldsymbol{x}\|_2.$$

How to design W?

Theorem 3.1. *Assume that we have MK vectors $\boldsymbol{w}_{i,m}$ of length ℓ in which each entry is a scaled i.i.d. (sub-)Gaussian random variable with zero mean and unit variance (the scaling factor is $1/\sqrt{M\ell}$). Let t be the stride length (where $n = (D-\ell)/t+1$) and $\mathbf{W}$ be a structured random matrix, which is the weight matrix of a single layer CNN with M channels and input length D. If*

$$\frac{M\ell^2}{D} \geq \frac{C}{\delta_k^2}\Big(k(\log(K)+\log(n))-\log(\epsilon)\Big)$$

for a positive constant C, then with probability $1-\epsilon$, the $MD \times Kn$ matrix $\mathbf{W}^T$ satisfies the model-RIP for model $\mathcal{M}_k$ with parameter δ_k.

We also note that the same analysis can be applied to the sum of two model-k-sparse signals, with changes in the constants (that we do not track here).

Corollary 3.2. *Random matrices with the CNN structure satisfy, with high probability, the model-RIP for $\mathcal{M}_k^2$.*