DCFNet: Deep Neural Network with Decomposed Convolutional Filters Supplementary Material

A. Proofs

In the proofs, some technical details are omitted for brevity and readability. The full proofs are left to the long version of the work.

Proof of Proposition 3.1. To prove (a), omitting (l) in $W^{(l)}$, and let $M = M_l$, $M' = M_{l-1}$, $B_{\lambda',\lambda} = \|W_{\lambda',\lambda}\|_1$. By definition of B_l, we have that

$$\begin{aligned} \sum_{\lambda' \in [M']} B_{\lambda',\lambda} \le B_l, \quad \forall \lambda \\ \sum_{\lambda \in [M]} B_{\lambda',\lambda} \le B_l \frac{M}{M'}, \quad \forall \lambda'. \end{aligned} \tag{A1}$$

We essentially use Schur's test, being more careful with the summation over λ'. We derive by Cauchy-Schwarz which is equivalent to Schur's test:

$$\begin{aligned}
&\|x^{(l)}[x_1] - x^{(l)}[x_2]\|^2 \cdot |\Omega| M \\
=& \sum_{\lambda \in [M]} \int \left| \sigma(\sum_{\lambda' \in [M']} \int x_1(u+v',\lambda') W_{\lambda',\lambda}(v') dv' + b(\lambda)) - \sigma(\sum_{\lambda' \in [M']} \int x_2(u+v',\lambda') W_{\lambda',\lambda}(v') dv' + b(\lambda)) \right|^2 du \\
\le& \sum_{\lambda \in [M]} \int \left| \sum_{\lambda' \in [M']} \int x_1(u+v',\lambda') W_{\lambda',\lambda}(v') dv' - \sum_{\lambda' \in [M']} \int x_2(u+v',\lambda') W_{\lambda',\lambda}(v') dv' \right|^2 du \\
=& \sum_{\lambda \in [M]} \int \left| \sum_{\lambda' \in [M']} \int (x_1 - x_2)(\tilde{v},\lambda') W_{\lambda',\lambda}(\tilde{v} - u) d\tilde{v} \right|^2 du \\
\le& \sum_{\lambda \in [M]} \int \left(\sum_{\lambda_1' \in [M']} \int |(x_1 - x_2)(v_1,\lambda_1')|^2 \left| W_{\lambda_1',\lambda}(v_1 - u) \right| dv_1 \right) \cdot \left(\sum_{\lambda_2' \in [M']} \|W_{\lambda_2',\lambda}\|_1 \right) du \\
\le& B_l \cdot \sum_{\lambda_1' \in [M']} \int |(x_1 - x_2)(v_1,\lambda_1')|^2 \left(\sum_{\lambda \in [M]} \|W_{\lambda_1',\lambda}\|_1 \right) dv_1 \\
\le& B_l \cdot B_l \frac{M}{M'} \cdot \|x_1 - x_2\|^2 |\Omega| M' = B_l^2 M \|x_1 - x_2\|^2 |\Omega|,
\end{aligned}$$

which means that

$$\|x^{(l)}[x_1] - x^{(l)}[x_2]\| \le B_l \|x_1 - x_2\|.$$

Thus $B_l \le 1$ implies (a).

To prove (b), we firstly verify that $x_0^{(l)}(\lambda)$ indeed is a constant over space for all λ and l. When $l = 0$, $x_0^{(0)}$ is all zero, so the claim is true. Suppose that the claim holds for $l-1$, then

$$x_0^{(l)}(u,\lambda) = \sigma\left(\sum_{\lambda'} \int x_0^{(l-1)}(\lambda') W_{\lambda',\lambda}^{(l)}(v') dv' + b^{(l)}(\lambda) \right)$$

which again does not depend on u. So we can write $x_0^{(l)}$ as $x_0^{(l)}(\lambda)$. Now by (a),

$$\|x_c^{(l)}\| = \|x^{(l)}[x^{(l-1)}] - x^{(l)}[x_0^{(l-1)}]\| \leq \|x^{(l-1)} - x_0^{(l-1)}\| = \|x_c^{(l-1)}\|,$$

which proves (b). □

Proof of Lemma 3.2. To illustrate the idea, we first prove the lemma in the one-dimensional case, i.e. $u \in \mathbb{R}$ instead of $\mathbb{R}^2$. We then extend to the 2D case. In the 1D case, the constant c_1 can be improved to be 2, and we only need $|\tau'|_\infty < \frac{1}{2}$. In the 2D case, we need $c_1 = 4$ as in the final claim.

To simply notation, we denote the mapping $x^{(l)}[x^{(l-1)}]$ as $y[x]$, $x_c^{(l-1)}$ by x_c, $M_{l-1} = M'$, $M_l = M$, and $W^{(l)}$ by W. Let $C_{\lambda',\lambda} = \int |v||\frac{d}{dv}W_{\lambda',\lambda}(v)|dv$, and $B_{\lambda',\lambda} = \int |W_{\lambda',\lambda}(v)|dv$, then (A1) holds, and the same relation holds for $C_{\lambda',\lambda}$ and C_l.

By definition,

$$D_\tau y[x](u,\lambda) = \sigma\left(\sum_{\lambda' \in [M']} \int x(\rho(u) + v', \lambda')W_{\lambda',\lambda}(v')dv' + b(\lambda)\right),$$

$$y[D_\tau x](u,\lambda) = \sigma\left(\sum_{\lambda' \in [M']} \int x(\rho(u + v'), \lambda')W_{\lambda',\lambda}(v')dv' + b(\lambda)\right).$$

Relaxing by removing σ as in the proof of Proposition 3.1, one can derive that

$$\|D_\tau y[x] - y[D_\tau x]\|^2 \cdot |\Omega| M \leq \|E_1 + E_2\|^2,$$

where

$$E_1(u,\lambda) = \sum_{\lambda' \in [M']} \int x_c(v,\lambda')(W_{\lambda',\lambda}(v - \rho(u)) - W_{\lambda',\lambda}(\rho^{-1}(v) - u))dv,$$

$$E_2(u,\lambda) = \sum_{\lambda' \in [M']} \int x_c(v,\lambda')W_{\lambda',\lambda}(\rho^{-1}(v) - u)(|(\rho^{-1})'(v)| - 1)dv.$$

Notice that x is replaced by x_c due to the fact that x and x_c differ by a constant field over space for each channel λ'. We bound $\|E_1\|$ and $\|E_2\|$ respectively.

For E_1, we introduce $k^{(1)}_{\lambda',\lambda}(v,u) = W_{\lambda',\lambda}(v - \rho(u)) - W_{\lambda',\lambda}(\rho^{-1}(v) - u)$, and re-write it as

$$E_1(u,\lambda) = \sum_{\lambda' \in [M']} \int x_c(v,\lambda')k^{(1)}_{\lambda',\lambda}(v,u)dv.$$

Applying Schur's test as in the proof of Proposition 3.1, one can show that

$$\|E_1\| \leq 2|\tau'|_\infty C_l \sqrt{M|\Omega|}\|x_c\|$$

as long as for all λ', λ,

$$\sup_u \int \left|k^{(1)}_{\lambda',\lambda}(v,u)\right| dv, \sup_v \int \left|k^{(1)}_{\lambda',\lambda}(v,u)\right| du \leq 2C_{\lambda',\lambda}|\tau'|_\infty. \tag{A2}$$

(A2) can be verified by 1D change of variable, and details omitted.

For E_2, we introduce $k^{(2)}_{\lambda',\lambda}(v,u) = W_{\lambda',\lambda}(\rho^{-1}(v) - u)(|(\rho^{-1})'(v)| - 1)$, and then we have that

$$\int |k^{(2)}_{\lambda',\lambda}(v,u)|du \leq |(\rho^{-1})'(v) - 1| \cdot \int |W_{\lambda',\lambda}(u)|du \leq 2|\tau'|_\infty B_{\lambda',\lambda}, \quad \forall v,$$

where we use $1-(\rho^{-1})'(t) = \frac{-\tau'(\rho^{-1}(t))}{1-\tau'(\rho^{-1}(t))}$ and $|\tau'| < \frac{1}{2}$ to obtain the factor 2. Meanwhile,

$$\int |k^{(2)}_{\lambda',\lambda}(v,u)|dv = \int |W_{\lambda',\lambda}(\tilde{v}-u)||1-|\rho'(\tilde{v})||d\tilde{v} \le |\tau'|_\infty B_{\lambda',\lambda}, \quad \forall u.$$

This gives that

$$\|E_2\| \le 2|\tau'|_\infty B_l \sqrt{M|\Omega|}\|x_c\|.$$

Putting together we have that

$$\sqrt{M|\Omega|}\|D_\tau y[x] - y[D_\tau x]\| \le \|E_1 + E_2\| \le \|E_1\| + \|E_2\| \le 2|\tau'|_\infty (C_l + B_l)\sqrt{M|\Omega|}\|x_c\|$$

which proves the claim in the 1D case.

The extension to the 2D case uses standard elementary techniques. The assumption $|\nabla\tau|_\infty < \frac{1}{5}$ is used to derive that $||J\rho|-1|$, $||J\rho^{-1}|-1| \le 4|\nabla\tau|_\infty$, and $|J\rho|$, $|J\rho^{-1}| \le 2$. In all the formula, $|(\rho^{-1})'(v)|$ is replaced by the Jacobian determinant $|J\rho^{-1}(v)|$. The integration in 1D is replaced by that along a segment in the 2D space. Details omitted. □

Proof of Prop. 3.3. Under these conditions, Proposition 3.1 applies. Let $c_1 = 4$. Introduce the notation

$$y_l = x^{(L)} \circ \cdots \circ D_\tau x^{(l)} \circ \cdots \circ x^{(0)}, \quad l = 0, \cdots, L$$

where $y_0 = x^{(L)}[D_\tau x^{(0)}]$, and $y_L = D_\tau x^{(L)}[x^{(0)}]$. The l.h.s equals $\|y_0 - y_L\|$, and we will bound it by $\|y_L - y_0\| \le \sum_{l=1}^L \|y_l - y_{l-1}\|$. For each $l = 1, \cdots, L$,

$$\begin{aligned}
\|y_l - y_{l-1}\| =& \|x^{(L)} \circ \cdots \circ D_\tau x^{(l)} \circ x^{(l-1)} \\
&- x^{(L)} \circ \cdots \circ x^{(l)} \circ D_\tau x^{(l-1)}\| \\
\le& \|D_\tau x^{(l)} \circ x^{(l-1)} - x^{(l)} \circ D_\tau x^{(l-1)}\| \\
\le& c_1(C_l + B_l)|\nabla\tau|_\infty \|x_c^{(l-1)}\| \\
\le& 2c_1|\nabla\tau|_\infty \|x_c^{(l-1)}\| \\
\le& 2c_1|\nabla\tau|_\infty \|x^{(0)}\|,
\end{aligned}$$

where the first inequality is by the nonexpansiveness of the $(l+1)$ to L-th layer, the second by Lemma 3.2, the third by (A2), and the last by Proposition 3.1 (b). Thus, $\sum_{l=1}^L \|y_l - y_{l-1}\| \le 2c_1 L|\nabla\tau|_\infty \|x^{(0)}\|$. □

Proof of Proposition 3.4. The technique is similar to that in the proof of Lemma 3.2. Let the constant on the r.h.s be denoted by c_2. In the 1D case, the constant c_2 can be improved to be 1. In the 2D case, $c_2 = 2$ as in the final claim. Details omitted. □

Proof of Lemma 3.5. The first claim is a classical result, and has a direct proof as $\int_{D(0)} |\nabla F|^2 = -\int_{D(0)} F\Delta F = \langle \sum_k a_k\psi_k, \sum_k a_k\mu_k\psi_k\rangle = \pi\sum_k a_k^2\mu_k$ by the orthogonality of ψ_k, as stated above in the text. By Cauchy-Schwarz, $\|\nabla F\|_1 \le \sqrt{\pi}\|\nabla F\|_2$. Putting together gives the second claim. □

Proof of Proposition 3.6. Omitting λ', λ, l, and let $j_l = j$, we write $W(u) = \sum_k a_k\psi_{j,k}(u)$. Rescaled to $D(0)$, we consider $w(u) = \sum_k a_k\psi_k(u)$, and one can verify that $\||v||\nabla W(v)|\|_1 = \||v||\nabla w(v)|\|_1$, and $\|W\|_1 = \|w\|_1$. Meanwhile, $\int_{D(0)} |v||\nabla w(v)|dv \le \int_{D(0)} |\nabla w(v)|dv$ by that $|v| \le 1$, and $\|w\|_1 \le \|\nabla w\|_1$ by Poincaré inequality, using the fact that w vanishes on the boundary of $D(0)$. Thus $\||v||\nabla w|\|_1, \|w\|_1 \le \|\nabla w\|_1$. The claim of the proposition follows by applying Lemma 3.5 to w. □

Proof of Theorem 3.8. Let $c_1 = 4$, $c_2 = 2$. The l.h.s. is bounded by $\|x^{(L)} - D_\tau x^{(L)}\| + \|D_\tau x^{(L)}[x^{(0)}] - x^{(L)}[D_\tau x^{(0)}]\|$. The second term is less than $2c_1 L|\nabla\tau|_\infty\|x^{(0)}\|$ by Theorem 3.7. To bound the first term, we apply Proposition 3.4, and notice that for all λ', λ, $\|\nabla W^{(L)}_{\lambda',\lambda}\|_1 \le 2^{-j_L}\pi\|a^{(L)}_{\lambda,\lambda}\|_{FB}$ (consider $W^{(L)}_{\lambda',\lambda}(u) = W(u) = \sum_k a_k\psi_{J,k}(u) = 2^{-2J}\sum_k a_k\psi_k(2^{-J}u)$, $J = j_L$, let $w(u) = \sum_k a_k\psi_k(u)$, then $W(u) = 2^{-2J}w(2^{-J}u)$, and $\|\nabla W\|_1 = 2^{-J}\|\nabla w\|_1$, where $\|\nabla w\|_1 \le \sqrt{\pi}\|a\|_{FB}$ by Lemma 3.5), and thus $D_L \le 2^{-j_L}A_L$. By (A2'), this gives that $\|D_\tau x^{(L)} - x^{(L)}\| \le c_2 2^{-j_L}|\tau|_\infty \|x_c^{(L-1)}\|$, and $\|x_c^{(L-1)}\| \le \|x^{(0)}\|$ by Proposition 3.1 (b). □

B. Experimental Details

The training of a Conv-2 DCF-FB network (Table 2) on MNIST dataset:

The network is trained using standard Stochastic Gradient Descent (SGD) with momentum 0.9 and batch size 100 for 100 epochs. L^2 regularization ("weightdecay") of 10^{-4} is used on the trainable parameters a's. The learning rate decreases from 10^{-2} to 10^{-4} over the 100 epochs. Batch normalization is used after each convolutional layer. The typical evolution of training and testing losses and errors over epochs are shown in Figure B.1.

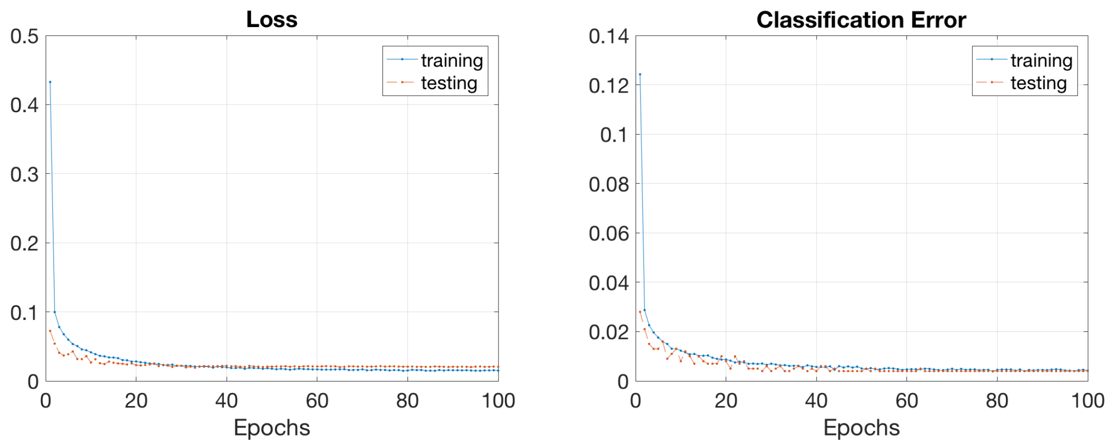

Figure B.1. The evolution of training and validation losses (left) and errors (right) over the epochs of a Conv-2 DCF-FB network trained on 50K MNIST using SGD.