Appendix

Proof of Equation (3): Note that

$$\begin{aligned} &F(x+e) \neq F(x) \\ \Leftrightarrow\ &\langle x+e, z\rangle\langle x, z\rangle < 0 \\ \Leftrightarrow\ &\langle e, z\rangle > |\langle x, z\rangle|. \end{aligned}$$

The left-hand side is clearly maximized for $e = \|e\|\frac{z}{\|z\|}$, leading to

$$\|e\|\|z\| > |\langle x, z\rangle|.$$

This proves the claim by taking the infimum over $\|e\|$.

Lemma 1. *Let F be a classifier with locally affine score function Ψ. Assume $l(x) \geq \rho(x)$. Then*

$$\rho(x) = \min_{j \neq i^*} \frac{\Psi^{i^*}(x) - \Psi^j(x)}{\|\nabla\Psi^{i^*}(x) - \nabla\Psi^j(x)\|}, \tag{8}$$

for $i^* := F(x)$ the predicted class at x.

Proof. As $l(x) \geq \rho(x)$, we can take the infimum in (1) over all perturbations in the local affine component, i.e. e with $\|e\| \leq l(x)$ only. This allows us to reformulate

$$\begin{aligned} &F(x+e) \neq F(x) \\ \Leftrightarrow\ &\exists j \neq i^* : \Psi^j(x+e) > \Psi^{i^*}(x+e) \\ \Leftrightarrow\ &\exists j \neq i^* : \langle\nabla\Psi^j(x) - \nabla\Psi^{i^*}(x), e\rangle > \Psi^{i^*}(x) - \Psi^j(x). \end{aligned}$$

The infimum over $\|e\|$ is achieved by choosing e as a multiple of $\nabla\Psi^j(x) - \nabla\Psi^{i^*}(x)$. A direct computation then finishes the proof. □

Proofs of Homogenization results

Lemma 3 (Euler's Homogeneous Function Theorem)**.** *Let $f : \mathbb{R}^m \to \mathbb{R}$ be a positive one-homogeneous function that is continuously differentiable on $\mathbb{R}^m \backslash \{0\}$. Then*

$$f(x) = \langle\nabla f(x), x\rangle$$

Proof. First note that

$$\begin{aligned} \partial_i f(ax) &= \lim_{t\to 0} \frac{f(ax + te_i) - f(ax)}{t} \\ &= \lim_{t\to 0} \frac{f(ax + ate_i) - f(ax)}{at} = \partial_i f(x). \end{aligned}$$

Hence

$$f(x) = \int_0^1 \langle\nabla f(tx), x\rangle\ \mathrm{dt} = \langle\nabla f(x), x\rangle$$

□

Lemma 2 (Linearized Robustness of Homogeneous Classifiers)**.** *Consider a classifier F with positive one-homogeneous score functions. Then*

$$\tilde{\rho}(x) = \alpha^\dagger(x). \tag{12}$$

Proof. Direct consequence of 3. □

Definition 5 (Neural Networks)**.** *Define the class of neural networks $\mathcal{N}$ to be any network built on learnable affine transforms (convolutional layers, dense layers) with linear weights Θ and biases b and ReLU or leaky ReLU activations. The network can include arbitrary skip-connections, batch-normalization layers and max or average pooling layers of arbitrary window size. This in particular includes many state-of-the-art classification networks.*

Lemma 4 (Homogeneous Networks)**.** *For fixed x, consider the logit $\Psi^i_{\Theta,b}(x)$ of a network $\Psi_{\Theta,b} \in \mathcal{N}$, where Θ denotes the linear weights and b the bias vector of the network. Then the function*

$$f : y \mapsto \Psi^i_{\Theta, b\frac{\|y\|}{\|x\|}}(y),$$

f is positive one-homogeneous and $f(x) = \Psi^i_{\Theta,b}(x)$.

Proof. Consider first a network consisting of a single layer with linear transform A and bias b with ReLU non-linearity. The associated network function is hence given by $\Psi_{A,b}(x) = (Ax + b)_+$. For this network, we compute for x fixed and any y and $a > 0$ as

$$\begin{aligned} f(ay) &= \left(A(ay) + b\frac{\|ay\|}{\|x\|}\right)_+ \\ &= \left(a \cdot Ay + a \cdot b\frac{\|y\|}{\|x\|}\right)_+ = af(y). \end{aligned}$$

A single layer is hence positive one-homogeneous. A function consisting of compositions of positive one-homogeneous functions is positive one-homogeneous itself as well, the function f associated to a network consisting of affine transforms and ReLU activations is positive one-homogeneous. All of the operations skip-connections, batch-normalization layers and max or average pooling are positive one-homogeneous as well, thus proving the claim. □

Theorem 1 (Homogeneous Decomposition of Neural Networks)**.** *Let $\Psi^i_{\Theta,b}$ be any logit of a neural network with ReLU activations (of class $\mathcal{N}$ in the appendix). Denote by Θ the linear filters and by b the bias terms of the network. Then*

$$\begin{aligned} \Psi^i_{\Theta,b}(x) &= \langle x, \nabla_x\Psi^i_{\Theta,b}(x)\rangle + \langle b, \nabla_b\Psi^i_{\Theta,b}(x)\rangle \\ &= \langle x, \nabla_x\Psi^i_{\Theta,b}(x)\rangle + \sum_k b_k\partial_{b_k}\Psi^i_{\Theta,b}(x). \end{aligned} \tag{13}$$

Proof. Let f be the functions associated with the network $\Psi^i_{\Theta,b}$ as in Lemma 4. Then by Lemma 3 we can compute the value of f at the point x via

$$f(x) = \langle x, \nabla_y f(y)|_{y=x}\rangle.$$

Note that by construction $f(x) = \Psi^i_{\Theta,b}(x)$. We compute the gradient of f at the point x explicitly as

$$\nabla_y f(y)|_{y=x} = \nabla_x \Psi^i_{\Theta,b}(x) + \frac{x}{\|x\|^2}\langle b, \nabla_b \Psi^i_{\Theta,b}(x)\rangle.$$

Combining these results shows

$$\begin{aligned} f(x) &= \langle x, \nabla_x \Psi^i_{\Theta,b}(x) + \frac{x}{\|x\|^2}\langle b, \nabla_b \Psi^i_{\Theta,b}(x)\rangle\rangle \\ &= \langle x, \nabla_x \Psi^i_{\Theta,b}(x)\rangle + \langle b, \nabla_b \Psi^i_{\Theta,b}(x)\rangle. \end{aligned}$$

□

Recall the notation $i^* = F(x)$ and j^* for the minimizer in j in (9).

Theorem 2. *Let $g := \nabla\Psi^{i^*}(x)$. Furthermore, let $g^\dagger := \nabla(\Psi^{i^*} - \Psi^{j^*})(x)$ and $\beta^\dagger := \beta^{i^*}(x) - \beta^{j^*}(x)$. Then*

$$\begin{aligned} \tilde{\rho}(x) &\leq \alpha^\dagger(x) + \frac{|\beta^\dagger|}{\|g^\dagger\|} && (14) \\ &\leq \alpha(x) + \|x\| \cdot \|\overline{g}^\dagger - \overline{g}\| + \frac{|\beta^\dagger|}{\|g^\dagger\|}. && (15) \end{aligned}$$

Proof. We have

$$\begin{aligned} \tilde{\rho}(x) &= \frac{\Psi^{i^*}(x) - \Psi^{j^*}(x)}{\|\nabla\Psi^{i^*}(x) - \nabla\Psi^{j^*}(x)\|} \\ &= \frac{\langle x, \nabla\Psi^{i^*}(x) - \nabla\Psi^{j^*}(x)\rangle + \beta^{i^*}(x) - \beta^{j^*}(x)}{\|\nabla\Psi^{i^*}(x) - \nabla\Psi^{j^*}(x)\|} \\ &= \left|\langle x, \overline{g}^\dagger\rangle + \frac{\beta^\dagger}{\|g^\dagger\|}\right| \leq \alpha^\dagger(x) + \frac{|b^\dagger|}{\|g^\dagger\|}, \end{aligned}$$

using the decomposition theorem and the triangle inequality. Further,

$$\begin{aligned} &\alpha^\dagger(x) + \frac{|b^\dagger|}{\|g^\dagger\|} \\ &= \left|\langle x, \overline{g}^\dagger\rangle\right| + \frac{|b^\dagger|}{\|g^\dagger\|} \\ &= \left|\langle x, \overline{g}^\dagger - \overline{g} + \overline{g}\rangle\right| + \frac{|b^\dagger|}{\|g^\dagger\|} \\ &\leq |\langle x, \overline{g}\rangle| + \left|\langle x, \overline{g}^\dagger - \overline{g}\rangle\right| + \frac{|b^\dagger|}{\|g^\dagger\|} \\ &\leq \alpha(x) + \|x\| \cdot \|\overline{g}^\dagger - \overline{g}\| + \frac{|b^\dagger|}{\|g^\dagger\|}, \end{aligned}$$

using the Cauchy-Schwarz inequality. □

Theorem 3. *Let $\xi := x + \frac{\beta^\dagger}{\|g^\dagger\|}\frac{g^\dagger}{\|g^\dagger\|}$ and $\gamma := \nabla\Psi^{i^*}(\xi)$, with $g^\dagger$ and $\beta^\dagger$ defined as in the previous theorem. Then*

$$\tilde{\rho}(x) \leq \frac{|\langle \xi, \gamma\rangle|}{\|\gamma\|} + \|\xi\| \cdot \|\overline{g}^\dagger - \overline{\gamma}\|, \tag{16}$$

and if additionally $F(x) = F(\xi)$, then

$$\tilde{\rho}(x) \leq \alpha(\xi) + \|\xi\| \cdot \|\overline{g}^\dagger - \overline{\gamma}\|.$$

Proof. We have

$$\begin{aligned} \tilde{\rho}(x) &= \frac{\langle x, g^\dagger\rangle + \beta^\dagger\langle \frac{g^\dagger}{\|g^\dagger\|^2}, g^\dagger\rangle}{\|g^\dagger\|} \\ &= \frac{\langle x + \frac{\beta^\dagger}{\|g^\dagger\|}\frac{g^\dagger}{\|g^\dagger\|}, g^\dagger\rangle}{\|g^\dagger\|} \\ &= \langle \xi, \overline{g}^\dagger\rangle = \langle \xi, \overline{g}^\dagger - \overline{g} + \overline{g}\rangle \\ &\leq |\langle \xi, \overline{\gamma}\rangle| + \|\xi\| \cdot \|\overline{g}^\dagger - \overline{\gamma}\|, \end{aligned}$$

using the Cauchy-Schwarz inequality in the same way as in the last theorem. □

MNIST Model Architecture

Here we describe the architecture that was used for the MNIST models.

Conv2D (3 × 3, 'same'), 32 feature maps, ReLU
Max Pooling (factor 2)
Conv2D (3 × 3, 'same'), 64 feature maps, ReLU
Max Pooling (factor 2)
Conv2D (3 × 3, 'same'), 128 feature maps, ReLU
Max Pooling (factor 2)
Dense Layer (128 neurons), ReLU
Dropout (0.5)
Softmax