Supplementary Material for "On the Number of Linear Regions of Convolutional Neural Networks"

Huan Xiong [1] **Lei Huang** [2] **Mengyang Yu** [2] **Li Liu** [2] **Fan Zhu** [2] **Ling Shao** [1 2]

1. Preliminary on Hyperplane Arrangements

In this section, we recall some basic knowledge on hyperplane arrangements (Zaslavsky, 1975; Stanley, 2004), which will be used in the proofs of theorems in this paper. An affine hyperplane in a Euclidean space $V \simeq \mathbb{R}^n$ is a subspace with the following form: $H = \{X \in V : \alpha \cdot X = b\}$, where " $\cdot$ " denotes the inner product, $\mathbf{0} \neq \alpha \in V$ is called the *norm vector* of H, and $b \in \mathbb{R}$. For example, when $V = \mathbb{R}^n$, an affine hyperplane has the following form: $\{(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n : \sum_{i=1}^n a_i x_i = b\}$ where $a_i, b \in \mathbb{R}$ and there exists some i with $a_i \neq 0$. A finite *hyperplane arrangement* $\mathcal{A}$ of a Euclidean space V is a finite set of affine hyperplanes in V. A *region* of an arrangement $\mathcal{A} = \{H_i \subset V : 1 \leq i \leq m\}$ is defined as a connected component of $V \setminus (\cup_{i=1}^m H_i)$, which is a connected component of the complement of the union of the hyperplanes in $\mathcal{A}$. Let $r(\mathcal{A})$ denote the number of regions for an arrangement $\mathcal{A}$. It is natural to ask: What is the maximal number of regions for an arrangement with m hyperplanes in $\mathbb{R}^n$? The following Zaslavsky's theorem answers this question.

Proposition 1 (Zaslavsky's Theorem (Zaslavsky, 1975; Stanley, 2004))**.** *Let $\mathcal{A} = \{H_i \subset V : 1 \leq i \leq m\}$ be an arrangement in $\mathbb{R}^n$. Then, the number of regions for the arrangement $\mathcal{A}$ satisfies*

$$r(\mathcal{A}) \leq \sum_{i=0}^{n} \binom{m}{i}. \tag{1}$$

Furthermore, the above equality holds iff $\mathcal{A}$ is in general position, i.e., (i) $\dim(\bigcap_{j=1}^k H_{i_j}) = n - k$ for any $k \leq n$ and $1 \leq i_1 < i_2 < \cdots < i_j \leq m$; (ii) $\bigcap_{j=1}^k H_{i_j} = \emptyset$ for any $k > n$ and $1 \leq i_1 < i_2 < \cdots < i_j \leq m$.

For example, if $n = 2$ then a set of lines is in general position if no two are parallel and no three meet at a point. In this case, the number of regions of an arrangement $\mathcal{A}$ with m lines in general position is equal to

$$r(\mathcal{A}) = \binom{m}{2} + m + 1. \tag{2}$$

For an arrangement $\mathcal{A}$ and some $H_0 \in \mathcal{A}$, we define

$$\mathcal{A}^{H_0} := \{H \cap H_0 : H \in \mathcal{A}, H \neq H_0,\ H \cap H_0 \neq \emptyset\}$$

to be the set of nonempty intersections of H_0 and other hyperplanes in $\mathcal{A}$. The following lemma gives a recursive method to compute $r(\mathcal{A})$.

Lemma 1 (Lemma 2.1 from (Stanley, 2004))**.** *Let $\mathcal{A}$ be an arrangement and $H_0 \in \mathcal{A}$. Then we have*

$$r(\mathcal{A}) = r(\mathcal{A} \setminus \{H_0\}) + r(\mathcal{A}^{H_0}).$$

Lemma 1 means that we can calculate the number of regions of an arrangement by induction.

Let $\#\mathcal{A}$ be the number of hyperplanes in $\mathcal{A}$ and rank($\mathcal{A}$) be the dimension of the space spanned by the normal vectors of the hyperplanes in $\mathcal{A}$. An arrangement $\mathcal{A}$ is called *central* if $\bigcap_{H \in \mathcal{A}} H \neq \emptyset$.

[1]Mohamed bin Zayed University of Artificial Intelligence, UAE [2]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. Correspondence to: Huan Xiong <huan.xiong@mbzuai.ac.ae>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Lemma 2 (Theorems 2.4 and 2.5 from (Stanley, 2004))**.** *Let $\mathcal{A}$ be an arrangement in an n-dimensional vector space. Then we have*

$$r(\mathcal{A}) = \sum_{\substack{\mathcal{B}\subseteq\mathcal{A}\\ \mathcal{B}\ \mathit{central}}} (-1)^{\#\mathcal{B}-\mathit{rank}(\mathcal{B})}.$$

2. Proofs of Results for One-Layer CNNs

Let $[n,m] := \{n, n+1, n+2\ldots, m\}$ be the set of integers from n to m and $[m] := [1,m] = \{1,2,\ldots,m\}$. We establish the following generalization of Zaslavsky's theorem, which is crucial in the proof of Theorem 2.

Proposition 2. *Let $V = \mathbb{R}^n$, $V_1, V_2, \ldots, V_m$ be m nonempty subspaces of V, and $n_1, n_2, \ldots, n_m \in \mathbb{N}$ be some nonnegative integers. Let $\mathcal{A} = \{H_{k,j} : 1 \le k \le m,\ \ 1 \le j \le n_k\}$ be an arrangement in $\mathbb{R}^n$ with $H_{k,j} = \{X \in V : \alpha_{k,j}\cdot X = b_{k,j}\}$ where $\mathbf{0} \ne \alpha_{k,j} \in V_k,\ b_{k,j} \in \mathbb{R}$. Then, the number of regions for the arrangement $\mathcal{A}$ satisfies*

$$r(\mathcal{A}) \le \sum_{(i_1,i_2,\ldots,i_m)\in K_{V;V_1,V_2,\ldots,V_m}} \prod_{k=1}^{m} \binom{n_k}{i_k}, \tag{3}$$

where

$$K_{V;V_1,V_2,\ldots,V_m} = \left\{(i_1,i_2,\ldots,i_m) : i_k \in \mathbb{N},\ \ \sum_{k\in J} i_k \le \dim\left(\sum_{k\in J} V_k\right) \forall J \subseteq [m]\right\}.$$

Furthermore, assume that the following two conditions hold for the arrangement $\mathcal{A}$:

(i) For each $(i_1, i_2, \ldots, i_m) \in K_{V;V_1,V_2,\ldots,V_m}$, any $\sum_{k=1}^m i_k$ vectors with i_k distinct vectors chosen from the set $\{\alpha_{k,j} : 1 \le j \le n_k\}$ are linear independent;

(ii) For each $(i_1, i_2, \ldots, i_m) \in \mathbb{N}^m \setminus K_{V;V_1,V_2,\ldots,V_m}$, the intersection of any $\sum_{k=1}^m i_k$ hyperplanes with i_k distinct hyperplanes chosen from the set $\{H_{k,j} : 1 \le j \le n_k\}$ are empty.

Then, the equality in (3) *holds:*

$$r(\mathcal{A}) = \sum_{(i_1,i_2,\ldots,i_m)\in K_{V;V_1,V_2,\ldots,V_m}} \prod_{k=1}^{m} \binom{n_k}{i_k}. \tag{4}$$

Proof. First, we will prove (3) by induction on $\sum_{k=1}^m n_k$. When $\sum_{k=1}^m n_k = 0$, both sides of (3) equals 1 since $\binom{0}{0} = 1$. When $\sum_{k=1}^m n_k = 1$, both sides equals 2 since $\binom{1}{0} + \binom{1}{1} = 2$. Suppose that the result is true for $\sum_{k=1}^m n_k \le N$ for some $N \ge 1$. Now consider the case $\sum_{k=1}^m n_k = N+1$. Without loss of generality, assume $n_1 \ge 1$. Then $H_{1,1} \in \mathcal{A}$. Notice that the translation $Y \to Y + Y_0$ for some $Y_0 \in \mathbb{R}^n$ (i.e., translate all points in $\mathbb{R}$ by a vector Y_0) doesn't change the number of regions in $\mathcal{A}$. Thus we can assume $b_{1,1} = 0$. Then $H_{1,1}$ becomes an $(n-1)$-dimensional subspace of V. Replace H_0 in Lemma 1 with $H_{1,1}$, we obtain

$$r(\mathcal{A}) = r(\mathcal{A}\setminus\{H_{1,1}\}) + r(\mathcal{A}^{H_{1,1}}). \tag{5}$$

By induction hypothesis, we have

$$r(\mathcal{A}\setminus\{H_{1,1}\}) \le \sum_{(i_1,i_2,\ldots,i_m)\in K_{V;V_1,V_2,\ldots,V_m}} \binom{n_1-1}{i_1} \prod_{k=2}^{m} \binom{n_k}{i_k} \tag{6}$$

and

$$r(\mathcal{A}^{H_{1,1}}) \le \sum_{(i_1,i_2,\ldots,i_m)\in K_{V\cap H_{1,1};V_1\cap H_{1,1},V_2\cap H_{1,1},\ldots,V_m\cap H_{1,1}}} \binom{n_1-1}{i_1} \prod_{k=2}^{m} \binom{n_k}{i_k}. \tag{7}$$

Let's consider (7) first. Since $H_{1,1}$ is the orthogonal complement of the linear subspace generated by $\alpha_{1,1}$, and $\mathbf{0} \ne \alpha_{1,1} \subset V_1$, we have

$$H_{1,1} + V_1 = V.$$

Let $V_k' = H_{1,1} \cap V_k$ for $1 \le k \le m$. Therefore, for each $J \subseteq [2, m]$, we have

$$\dim\left(H_{1,1} \cap \left(V_1 + \sum_{k\in J} V_k\right)\right) = \dim(H_{1,1}) + \dim\left(V_1 + \sum_{k\in J} V_k\right) - \dim(V) = \dim\left(V_1 + \sum_{k\in J} V_k\right) - 1 \tag{8}$$

and thus

$$\dim\left(V_1' + \sum_{k\in J} V_k'\right) = \dim\left(V_1 + \sum_{k\in J} V_k\right) - 1. \tag{9}$$

On the other hand, it is trivial that

$$\dim\left(\sum_{k\in J} V_k'\right) \le \dim\left(\sum_{k\in J} V_k\right) \tag{10}$$

for any $J \subseteq [2, m]$. Therefore, by (7) we derive

$$\begin{aligned}
r(\mathcal{A}^{H_{1,1}}) &\le \sum_{(i_1,i_2,\ldots,i_m)\in K_{H_{1,1};V_1',V_2',\ldots,V_m'}} \binom{n_1-1}{i_1}\prod_{k=2}^{m}\binom{n_k}{i_k} \\
&\le \sum_{\substack{i_1-1+\sum_{k\in J} i_k \le \dim\left(V_1'+\sum_{k\in J} V_k'\right)\ \forall J\subseteq[2,m] \\ \sum_{k\in J} i_k \le \dim\left(\sum_{k\in J} V_k'\right)\ \forall J\subseteq[2,m]}} \binom{n_1-1}{i_1-1}\prod_{k=2}^{m}\binom{n_k}{i_k} \\
&\le \sum_{\substack{i_1+\sum_{k\in J} i_k \le \dim\left(V_1+\sum_{k\in J} V_k\right)\ \forall J\subseteq[2,m] \\ \sum_{k\in J} i_k \le \dim\left(\sum_{k\in J} V_k\right)\ \forall J\subseteq[2,m]}} \binom{n_1-1}{i_1-1}\prod_{k=2}^{m}\binom{n_k}{i_k} \\
&= \sum_{(i_1,i_2,\ldots,i_m)\in K_{V;V_1,V_2,\ldots,V_m}} \binom{n_1-1}{i_1-1}\prod_{k=2}^{m}\binom{n_k}{i_k}.
\end{aligned} \tag{11}$$

Put (5), (6) and (11) together, we obtain

$$\begin{aligned}
r(\mathcal{A}) &\le \sum_{(i_1,i_2,\ldots,i_m)\in K_{V;V_1,V_2,\ldots,V_m}} \left(\binom{n_1-1}{i_1}\prod_{k=2}^{m}\binom{n_k}{i_k} + \binom{n_1-1}{i_1-1}\prod_{k=2}^{m}\binom{n_k}{i_k}\right) \\
&= \sum_{(i_1,i_2,\ldots,i_m)\in K_{V;V_1,V_2,\ldots,V_m}} \prod_{k=1}^{m}\binom{n_k}{i_k},
\end{aligned} \tag{12}$$

which competes the proof of (3).

Furthermore, assume that the arrangement $\mathcal{A}$ satisfies the condition (i) and (ii). Then, the central sub-arrangements of $\mathcal{A}$ are exactly the sub-arrangements $\mathcal{B}$ consisting of $\sum_{k=1}^{m} i_k$ hyperplanes with i_k distinct hyperplanes chosen from the set $\{H_{k,j} : 1 \le j \le n_k\}$, where $(i_1, i_2, \ldots, i_m) \in K_{V;V_1,V_2,\ldots,V_m}$. In this case, $\#\mathcal{B} = \mathrm{rank}(\mathcal{B}) = \sum_{k=1}^{m} i_k$. Also, for any given $(i_1, i_2, \ldots, i_m) \in K_{V;V_1,V_2,\ldots,V_m}$, we have $\binom{n_k}{i_k}$ choices to pick i_k hyperplanes from each $\{\alpha_{k,i} : 1 \le i \le n_k\}$. Therefore, by Lemma 2 we obtain

$$r(\mathcal{A}) = \sum_{\substack{\mathcal{B}\subseteq\mathcal{A} \\ \mathcal{B}\text{ central}}} (-1)^{\#\mathcal{B}-\mathrm{rank}(\mathcal{B})} = \sum_{\substack{\mathcal{B}\subseteq\mathcal{A} \\ \mathcal{B}\text{ central}}} 1 = \sum_{(i_1,i_2,\ldots,i_m)\in K_{V;V_1,V_2,\ldots,V_m}} \prod_{k=1}^{m}\binom{n_k}{i_k}.$$

$\square$

To prove Theorem 2, we need the following lemmas on picking distinct elements from the union of certain sets.

Lemma 3. *Let $S_1, S_2, \dots S_m$ be m finite sets, and $a_1, a_2, \dots a_m$ be some nonnegative integers such that for any $I \subseteq [m]$,*

$$\sum_{i\in I} a_i \leq \# \bigcup_{i\in I} S_i. \tag{13}$$

Then, we can take a_i elements from each S_i such that these $\sum_{i=1}^m a_i$ elements are distinct.

Proof. We will prove this lemma by induction on m. When $m = 1$, the claim is trivial. Now assume that the lemma holds for any $1 \leq m < n$ and consider the case $m = n$. Without loss of generality, we assume that there exists some $\emptyset \neq I \subseteq [n]$ such that (otherwise we can always increase some a_i to make the following equality holds for some I)

$$\sum_{i\in I} a_i = \# \bigcup_{i\in I} S_i. \tag{14}$$

The proof is divided into two cases.

Case (1): There exists some I satisfying (14) with $\emptyset \neq I \neq [n]$. In this case, we can assume that $I = [r]$ for some $1 \leq r \leq n-1$ by symmetry, i.e.,

$$\sum_{i=1}^{r} a_i = \# \bigcup_{i=1}^{r} S_i. \tag{15}$$

Let

$$S'_j = S_{j+r} \setminus \bigcup_{i=1}^{r} S_i, \qquad 1 \leq j \leq n-r.$$

Then $\left(\bigcup_{j\in J} S'_j\right) \cap \left(\bigcup_{i=1}^r S_i\right) = \emptyset$. Therefore, for any $J \subseteq [n-r]$, we have

$$\# \bigcup_{j\in J} S'_j = \# \left(\bigcup_{j\in J} S'_j \cup \bigcup_{i=1}^{r} S_i\right) - \# \bigcup_{i=1}^{r} S_i = \# \left(\bigcup_{j\in J} S_{j+r} \cup \bigcup_{i=1}^{r} S_i\right) - \# \bigcup_{i=1}^{r} S_i. \tag{16}$$

By (13) and (15) the above equality becomes

$$\# \bigcup_{j\in J} S'_j \geq \left(\sum_{j\in J} a_{j+r} + \sum_{i=1}^{r} a_i\right) - \sum_{i=1}^{r} a_i = \sum_{j\in J} a_{r+j}. \tag{17}$$

Since $1 \leq \#I \leq n-1$, by induction we can pick a_i elements from each S_i for $1 \leq i \leq r$, and a_{r+j} elements from each S_{j+r} for $1 \leq j \leq n-r$ such that these $\sum_{i=1}^n a_i$ elements are distinct. Thus the claim holds.

Case (2): The only I satisfying (14) is $I = [n]$. Then $\#S_1 > a_1$ and thus $S_1 \cap \bigcup_{i=2}^n S_i \neq \emptyset$ (otherwise $\sum_{i=1}^n a_i = \# \bigcup_{i=1}^n S_i = \#S_1 + \# \bigcup_{i=2}^n S_i > \sum_{i=1}^n a_i$, a contradiction). Let $x \in S_1 \cap \bigcup_{i=2}^n S_i$ and

$$S'_j = \begin{cases} S_j, & 2 \leq j \leq n; \\ S_j \setminus \{x\}, & j = 1. \end{cases}$$

Then $\{S'_j : 1 \leq j \leq n\}$ still satisfies (13). But $\sum_{i=1}^n \#S'_i < \sum_{i=1}^n \#S_i$. Then $\{S'_j : 1 \leq j \leq n\}$ either satisfies Case (1), which leads to a solution; or still in Case (2), which we can continue the process until Case (i) satisfies. This completes the proof. $\square$

Lemma 4. *Let $S_1, S_2, \dots S_m$ be m finite sets. Then, there exist some $a_1, a_2, \dots a_m \in \mathbb{N}$ such that*

$$\sum_{i=1}^{m} a_i = \# \bigcup_{i=1}^{m} S_i, \tag{18}$$

and for any $I \subseteq [m]$,

$$\sum_{i\in I} a_i \leq \# \bigcup_{i\in I} S_i. \tag{19}$$

Proof. We will prove it by Induction on m. The claim is trivial when $m = 1$. Now assume that $m \geq 2$ and the result is true for $m - 1$. Therefore, we can pick some $a_1, a_2, \ldots a_{m-1} \in \mathbb{N}$ such that

$$\sum_{i=1}^{m-1} a_i = \# \bigcup_{i=1}^{m-1} S_i, \tag{20}$$

and for any $I \subseteq [m-1]$,

$$\sum_{i \in I} a_i \leq \# \bigcup_{i \in I} S_i. \tag{21}$$

Furthermore, let $a_m = \# \left(S_m \setminus \bigcup_{i=1}^{m-1} S_i \right)$. Then, for any $I \subseteq [m-1]$, we have

$$a_m + \sum_{i \in I} a_i \leq \# \bigcup_{i \in I} S_i + \# \left(S_m \setminus \bigcup_{i=1}^{m-1} S_i \right) \leq \# \bigcup_{i \in I \cup \{m\}} S_i. \tag{22}$$

Also,

$$\sum_{i=1}^{m} a_i = \# \bigcup_{i=1}^{m-1} S_i + \# \left(S_m \setminus \bigcup_{i=1}^{m-1} S_i \right) = \# \bigcup_{i=1}^{m} S_i. \tag{23}$$

Then the claim is also true for m. $\square$

We also need the following lemmas on measure zero subsets of Euclidean spaces with respect to Lebesgue measure.

Lemma 5. *Let $V \cong \mathbb{R}^n$ be a vector space. Then $S = \{(v_1, v_2, \ldots, v_n) \in V^n : v_1, v_2, \ldots, v_n$ are linear dependent$\}$ is a measure zero subset of V^n, with respect to Lebesgue measure.*

Proof. Without loss of generality, assume $V = \mathbb{R}^n$. Let the i-th vector be $v_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,n})$. Then $v_1, v_2, \ldots, v_n$ are linear dependent iff

$$det((x_{i,j})_{n \times n}) = 0,$$

whose left hand side is a non-zero polynomial of all $x_{i,j}$. It is easy to see that the solution of this polynomial has co-dimension 1 in $\mathbb{R}^{n \times n}$, thus S is a measure zero set. $\square$

Lemma 6. *Let $m > n$ be two given positive integers, $A = (a_{ij})_{m \times n} \in \mathbb{R}^{m \times n}$ and $C = (c_1, c_2, \ldots, c_m) \in \mathbb{R}^m$. Let S be the set of $(A, C) \in \mathbb{R}^{m(n+1)}$ such that*

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = c_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n = c_2 \\ \qquad\qquad \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = c_m \end{cases}$$

has solutions for $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$. Then S is a measure zero subset of $\mathbb{R}^{m(n+1)}$, with respect to Lebesgue measure.

Proof. By Lemma 5, the augmented matrix (A, C) has the rank $(n+1)$ except for a measure zero subset of $\mathbb{R}^{m(n+1)}$. On the other hand, the rank of the matrix A is at most n. Therefore, the rank of the augmented matrix (A, C) is larger than the rank of A except for a measure zero subset of $\mathbb{R}^{m(n+1)}$, thus by Rouché-Capelli Theorem (Shafarevich & Remizov, 2012) we obtain that (6) has no solutions except for a measure zero set of $\mathbb{R}^{m(n+1)}$. $\square$

Lemma 3 implies the following results when we choose a basis of a linear space properly.

Lemma 7. *Let $V \cong \mathbb{R}^n$ be a vector space and V_i $(1 \le i \le m)$ be m subspaces of V. Suppose that some non-negative integers a_i $(1 \le i \le m)$ satisfy*

$$\sum_{i \in I} a_i \le dim(\sum_{i \in I} V_i)$$

for each $I \subseteq [m]$. Then we obtain the following result.

(i) We can pick a_i vectors from V_i for $1 \le i \le m$ such that these $\sum_{1 \le i \le m} a_i$ vectors are linear independent.

(ii) $\sum_{1 \le i \le m} a_i$ vectors with a_i vectors from V_i for $1 \le i \le m$ such that they are linear dependent, forms a measure zero set in $\prod_{i=1}^m V_i^{a_i}$, with respect to Lebesgue measure.

Proof. (i) By linear algebra, we can construct a basis $v_1, v_2, \ldots, v_n$ of V such that each V_i has a basis which is a subset of $v_1, v_2, \ldots, v_n$. Then, by Lemma 3 this claim holds.

(ii) Let $n' = \sum_{1 \le i \le m} a_i$. By (i) there exist n' linear independent vectors $v_1, v_2, \ldots, v_{n'}$ with a_i vectors from V_i for $1 \le i \le m$. Let V_i' be the vector spaces generated by such a_i vectors in V_i. For any n' linear dependent vectors $v_1', v_2', \ldots, v_{n'}'$ with a_i vectors from V_i for $1 \le i \le m$, their projections $v_1'', v_2'', \ldots, v_{n'}''$ onto $\prod_{i=1}^m V_i'$ are also linear dependent. Suppose that $v_k'' = \sum_{j=1}^{n'} y_{k,j} v_j$ for $1 \le k \le n'$. If v_k' are chosen from V_{i_1}, such that $v_j \notin V_{i_1}$, we set $y_{k,j} = 0$. Otherwise, we set $y_{k,j} = y_{k,j}'$. Therefore, $\#\{y_{k,j}'\}$ equals the dimension of the projection of $\prod_{i=1}^m V_i^{a_i}$ onto $\prod_{i=1}^m V_i'$. Also, $v_1'', v_2'', \ldots, v_{n'}''$ are linear dependent iff

$$\det\left((y_{k,j})_{n' \times n'}\right) = 0.$$

Since $v_1, v_2, \ldots, v_{n'}$ are linear independent, the left hand side $\det\left((y_{k,j})_{n' \times n'}\right)$ must be a non-zero polynomial of some $y_{k,j}'$. Therefore, the solution of this polynomial forms a measure zero set in $\mathbb{R}^{\#\{y_{k,j}'\}}$ due to the zero measurability of the solutions of non-zero polynomial in Euclidean spaces (see (Lojasiewicz, 1964)). Thus such $\sum_{1 \le i \le m} a_i$ vectors forms a measure zero set in $\prod_{i=1}^m V_i^{a_i}$, with respect to Lebesgue measure. □

Now we are ready to prove Theorem 2.

Proof of Theorem 2. By Definition 1, the number of linear regions of $\mathcal{N}$ at θ is equal to the number of regions of the hyperplane arrangement

$$\mathcal{A}_{\mathcal{N},\theta} := \{H_{i,j,k}(X^0; \theta) : 1 \le i \le n_1^{(1)},\ 1 \le j \le n_1^{(2)}, 1 \le k \le d_1\},$$

where $H_{i,j,k}(X^0;\theta)$ is the hyperplane determined by $Z_{i,j,k}^1(X^0;\theta) = 0$ (the expression of $Z_{i,j,k}^1(X^0;\theta)$ is given in (2)). Recall that $X^0 = (X_{a,b,c}^0)_{n_0^{(1)} \times n_0^{(2)} \times d_0}$. Then $H_{i,j,k}(X^0;\theta)$ can be written as

$$\langle \alpha_{i,j,k}, X^0 \rangle_F + B^{1,k} = 0,$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius inner product, $\alpha_{i,j,k}$ is an $n_0^{(1)} \times n_0^{(2)} \times d_0$ dimensional tensor, whose $(a + (i-1)s_1, b + (j-1)s_1, c)$-th element is $W_{a,b,c}^{1,k}$ for all $1 \le a \le f_1^{(1)},\ 1 \le b \le f_1^{(2)},\ 1 \le c \le d_0$; and 0 otherwise. Let

$$V_{i,j} = \{\beta \in \mathbb{R}^{n_0^{(1)} \times n_0^{(2)} \times d_0} : \beta_{a',b',c'} = 0\ \ \forall (a',b',c') \ne (a + (i-1)s_1, b + (j-1)s_1, c)\}$$

be the subspace of $\mathbb{R}^{n_0^{(1)} \times n_0^{(2)} \times d_0}$ generated by $n_0^{(1)} \times n_0^{(2)} \times d_0$ dimensional tensors whose $(a+(i-1)s_1, b+(j-1)s_1, c)$-th element ranges over $\mathbb{R}$ for all $1 \le a \le f_1^{(1)},\ 1 \le b \le f_1^{(2)},\ 1 \le c \le d_0$; and 0 otherwise. Then $\alpha_{i,j,k} \in V_{i,j}$ for $1 \le k \le d_1$. By Proposition 2, we obtain

$$R_{\mathcal{N},\theta} = r(\mathcal{A}_{\mathcal{N},\theta}) \le \sum_{(t_{i,j})_{(i,j) \in I_{\mathcal{N}}} \in K_{V;(V_{i,j})_{(i,j) \in I_{\mathcal{N}}}}} \prod_{k=1}^{m} \binom{d_1}{t_{i,j}}, \tag{24}$$

where

$$K_{V;(V_{i,j})_{(i,j)\in I_{\mathcal{N}}}} = \{(t_{i,j})_{(i,j)\in I_{\mathcal{N}}} : \sum_{(i,j)\in J} t_{i,j} \le \dim\left(\sum_{(i,j)\in J} V_{i,j}\right) \forall J \subseteq I_{\mathcal{N}}\}$$
$$= \{(t_{i,j})_{(i,j)\in I_{\mathcal{N}}} : t_{i,j}\in\mathbb{N}, \quad \sum_{(i,j)\in J} t_{i,j} \le \#\cup_{(i,j)\in J} S_{i,j} \;\; \forall J \subseteq I_{\mathcal{N}}\},$$

which gives an upper bound for $R_{\mathcal{N},\theta}$ and $R_{\mathcal{N}}$. Next we will show that this upper bound can be reached except for a measure zero set in $\mathbb{R}^{\#weights+\#bias}$ with respect to Lebesgue measure. By Lemmas 6 and 7, when θ ranges over $\mathbb{R}^{\#weights+\#bias}$, the set of θ such that $A_{\mathcal{N},\theta}$ satisfies the conditions (i) and (ii) of Proposition 2 (replace $\{i_k : 1\le k\le m\}$ by $\{t_{i,j} : (i,j)\in I_{\mathcal{N}}\}$, and $\{V_k : 1\le k\le m\}$ by $\{V_{i,j} : (i,j)\in I_{\mathcal{N}}\}$), forms a complement of a measure zero set in $\mathbb{R}^{\#weights+\#bias}$, with respect to Lebesgue measure. Then, for such parameters θ, by Proposition 2 we derive the equality holds for (24), which implies that the maximal number $R_{\mathcal{N}}$ of linear regions of $\mathcal{N}$ is equal to

$$R_{\mathcal{N}} = \sum_{(t_{i,j})_{(i,j)\in I_{\mathcal{N}}}\in K_{\mathcal{N}}} \prod_{(i,j)\in I} \binom{d_1}{t_{i,j}},$$

and the right hand side of the above equality also equals the expectation of the number $R_{\mathcal{N},\theta}$ of linear regions of $\mathcal{N}$ with respect to the distribution μ of weights and biases. $\square$

The following result gives a simple example for Theorem 2.

Corollary 1. *Let $\mathcal{N}$ be a one-layer ReLU CNN with input dimension $1\times n\times 1$. Assume there are d_1 filters with dimension $1\times 2\times 1$ and stride $s=1$. Thus the hidden layer dimension is $1\times(n-1)\times d_1$. When n is fixed, we have*

$$R_{\mathcal{N}} = \frac{(n-1)}{2} d_1^n + \mathcal{O}(d_1^{n-1}). \tag{25}$$

Proof. By Theorem 2, we obtain

$$R_{\mathcal{N}} = \sum_{(t_{i,j})_{(i,j)\in I}\in K_{\mathcal{N}}} \prod_{(i,j)\in I} \binom{d_1}{t_{i,j}}. \tag{26}$$

Furthermore, when n is fixed, $R_{\mathcal{N}}$ is a polynomial of d_1 with degree n by Lemma 3 in the main paper. To calculate the coefficient of the leading term d_1^n of this polynomial, we need to determine all $(t_{i,j})_{(i,j)\in I_{\mathcal{N}}}\in K_{\mathcal{N}}$ with $\sum_{(i,j)\in I_{\mathcal{N}}} t_{i,j} = n$. First, since $n_1^{(1)}=1$ and $n_1^{(2)}=n-1$, it is easy to see that $I_{\mathcal{N}}=\{(1,j):\; 1\le j\le n-1\}$ and $S_{1,j}=\{(1,j,1),(1,j+1,1)\}$ for each $1\le j\le n-1$. Therefore,

$$K_{\mathcal{N}} = \{(t_{1,j})_{1\le j\le n-1}: \;\; t_{1,j}\in\mathbb{N}, \;\; \sum_{j\in J} t_{1,j} \le \#\cup_{(1,j)\in J} S_{1,j} \;\; \forall J\subseteq [n-1]\}. \tag{27}$$

Then, there are $n-1$ vectors $(t_{1,j})_{1\le j\le n-1}\in K_{\mathcal{N}}$ satisfying $\sum_{j=1}^{n-1} t_{1,j} = n$: $(2,1,1,\ldots,1)$, $(1,2,1,\ldots,1)$, $(1,1,2,1,\ldots,1)$, $\ldots$, $(1,1,1,\ldots,1,2)$. Therefore, the leading term in $R_{\mathcal{N}}$ equals

$$(n-1)\binom{d_1}{2} d_1^{n-2} = \frac{(n-1)}{2} d_1^n - \frac{(n-1)}{2} d_1^{n-1}$$

and thus

$$R_{\mathcal{N}} = \frac{(n-1)}{2} d_1^n + \mathcal{O}(d_1^{n-1}). \tag{28}$$

This completes the proof. $\square$

Next, we prove Lemma 3 and Theorem 3 in the main paper.

Proof of Lemma 3 *in the main paper.* Directly replace $\{a_i : 1 \le i \le m\}$ by $\{t_{i,j} : (i,j) \in I_{\mathcal{N}}\}$, and $\{S_i : 1 \le i \le m\}$ by $\{S_{i,j} : (i,j) \in I_{\mathcal{N}}\}$ in Lemma 4, we derive the result. □

Proof of Theorem 3. It is easy to see that $\binom{d_1}{t_{i,j}} = \Theta(d_1^{t_{i,j}})$ when d_1 tends to infinity. Then, by Eq. (4) and Lemma 3 in the main paper, we have

$$R_{\mathcal{N}} = \Theta(d_1^{\#\cup_{(i,j)\in I_{\mathcal{N}}} S_{i,j}}). \tag{29}$$

Furthermore, if all input neurons have been involved in the convolutional calculation, we have

$$\cup_{(i,j)\in I_{\mathcal{N}}} S_{i,j} = \{(a,b,c) : 1 \le a \le n_0^{(1)},\ 1 \le b \le n_0^{(2)},\ 1 \le c \le d_0\} \tag{30}$$

and thus

$$R_{\mathcal{N}} = \Theta(d_1^{n_0^{(1)} \times n_0^{(2)} \times d_0}).$$

□

3. Proofs of Results for Multi-Layer CNNs

In this section, we prove Theorem 5 on multi-layer ReLU CNNs.

Proof of Theorem 4. Assume that the parameters W and B for such two convolutional layers are the same as defined in Section 2. Let $l = 1, 2$ in (2) in the main paper and $X^l_{i,j,k} = Z^l_{i,j,k}(X^0; \theta)$, we obtain

$$X^1_{i,j,k} = \sum_{a=1}^{f_1^{(1)}} \sum_{b=1}^{f_1^{(2)}} \sum_{c=1}^{d_0} W^{1,k}_{a,b,c} X^0_{a+(i-1)s_1, b+(j-1)s_1, c} + B^{1,k} \tag{31}$$

and

$$X^2_{i,j,k} = \sum_{a=1}^{f_2^{(1)}} \sum_{b=1}^{f_2^{(2)}} \sum_{c=1}^{d_1} W^{2,k}_{a,b,c} X^1_{a+(i-1)s_2, b+(j-1)s_2, c} + B^{2,k}. \tag{32}$$

Substitute (31) into (32), we derive

$$X^2_{i,j,k} = \sum_{a'=1}^{f_2^{(1)}} \sum_{b'=1}^{f_2^{(2)}} \sum_{c'=1}^{d_1} \sum_{a=1}^{f_1^{(1)}} \sum_{b=1}^{f_1^{(2)}} \sum_{c=1}^{d_0} W^{2,k}_{a',b',c'} W^{1,c'}_{a,b,c} X^0_{a+(a'+(i-1)s_2-1)s_1, b+(b'+(j-1)s_2-1)s_1, c} + const \tag{33}$$

$$= \sum_{a'=1}^{f_2^{(1)}} \sum_{b'=1}^{f_2^{(2)}} \sum_{c'=1}^{d_1} \sum_{a=1}^{f_1^{(1)}} \sum_{b=1}^{f_1^{(2)}} \sum_{c=1}^{d_0} W^{2,k}_{a',b',c'} W^{1,c'}_{a,b,c} X^0_{a+(a'-1)s_1+(i-1)s_1s_2, b+(b'-1)s_1+(j-1)s_1s_2, c} + const. \tag{34}$$

Note that $1 \le a + (a'-1)s_1 \le f_1^{(1)} + (f_2^{(1)} - 1)s_1$ and $1 \le b + (b'-1)s_1 \le f_1^{(2)} + (f_2^{(2)} - 1)s_1$. Then (33) becomes

$$X^2_{i,j,k} = \sum_{a=1}^{f_1^{(1)}+(f_2^{(1)}-1)s_1} \sum_{b=1}^{f_1^{(2)}+(f_2^{(2)}-1)s_1} \sum_{c=1}^{d_0} {W'}^k_{a,b,c} X^0_{a+(i-1)s_2, b+(j-1)s_2, c} + const \tag{35}$$

where ${W'}^k_{a,b,c}$ are some constants. Therefore, $\mathcal{N}$ is realized as a ReLU CNN with one hidden convolutional layer such that its d_2 filters has size $(f_1^{(1)} + (f_2^{(1)} - 1)s_1) \times (f_1^{(2)} + (f_2^{(2)} - 1)s_1) \times d_0$ and stride $s_1 s_2$, which completes the proof. □

Proof of Theorem 5. (i) The basic idea is to map many regions of the input space of each layer to the same set, thus identify many regions of space.

The $L = 1$ case is guaranteed by Theorem 2. Next, we consider the case $L \geq 2$. Let $p = \lfloor d_1/d_0 \rfloor$. We set

$$W^{1,k}_{a,b,c} = \begin{cases} 1, \text{ if } a = b = 1, k = (c-1)p+1, \ \ 1 \leq c \leq d_0; \\ 2, \text{ if } a = b = 1, \ \ (c-1)p+2 \leq k \leq cp, 1 \leq c \leq d_0; \\ 0, \text{ otherwise} \end{cases} \tag{36}$$

and

$$B^{1,k} = \begin{cases} -(k-(c-1)p-1), \text{ if } (c-1)p+1 \leq k \leq cp \text{ for some } 1 \leq c \leq d_0; \\ 0, \text{ otherwise.} \end{cases} \tag{37}$$

Therefore, by (2) in the main paper we obtain

$$Z^1_{i,j,k}(X^0;\theta) = \begin{cases} X^0_{1+(i-1)s_1, 1+(j-1)s_1, c}, \text{ if } k = (c-1)p+1 \text{ for some } 1 \leq c \leq d_0; \\ 2X^0_{1+(i-1)s_1, 1+(j-1)s_1, c} - (k-(c-1)p-1), \text{ if } (c-1)p+2 \leq k \leq cp \text{ for some } 1 \leq c \leq d_0; \\ 0, \text{ otherwise.} \end{cases} \tag{38}$$

When $W^{1,k}_{a,b,c}$ and $B^{1,k}$ are given as in (36) and (37), the map

$$X^1_{i,j,k} = \max\{0, Z^1_{i,j,k}(X^0;\theta)\} \tag{39}$$

determines a function

$$X^1 = \Phi_1(X^0) \tag{40}$$

from $\mathbb{R}^{n_0^{(1)} \times n_0^{(2)} \times d_0}$ to $\mathbb{R}^{n_1^{(1)} \times n_1^{(2)} \times d_1}$.

For each $i, j \in \mathbb{N}^+$, let

$$\psi_i(x) = \begin{cases} \max\{0, x\}, \text{ if } i = 1; \\ \max\{0, 2x-(i-1)\}, \text{ if } i \geq 2 \end{cases} \tag{41}$$

and

$$\phi_j(x) = \sum_{i=1}^{j} (-1)^{i+1} \psi_i(jx). \tag{42}$$

Then it is easy to check that

$$\phi_j(x) = \begin{cases} 0, \text{ if } x \leq 0; \\ jx - i, \text{ if } \frac{i}{j} \leq x \leq \frac{2i+1}{2j} \leq \frac{1}{2} \text{ where } i \in \mathbb{N}; \\ i - jx, \text{ if } \frac{2i-1}{2j} \leq x \leq \frac{i}{j} \leq \frac{1}{2} \text{ where } i \in \mathbb{N}^+, \end{cases} \tag{43}$$

which means that ϕ_j is an affine function when restricted to each interval $[0, \frac{1}{2j}], [\frac{1}{2j}, \frac{2}{2j}], \ldots, [\frac{j-1}{2j}, \frac{1}{2}]$ and furthermore $\phi_j([0, \frac{1}{2j}]) = \phi_j([\frac{1}{2j}, \frac{2}{2j}]) = \cdots = \phi_j([\frac{j-1}{2j}, \frac{j}{2j}]) = [0, \frac{1}{2}]$ (i.e., $\phi_j(x)$ sends j distinct intervals $[0, \frac{1}{2j}], [\frac{1}{2j}, \frac{2}{2j}], \ldots, [\frac{j-1}{2j}, \frac{1}{2}]$ to the same interval $[0, \frac{1}{2}]$).

Next, we define an intermediate convolutional layer (without activation functions) from

$$X^1 = (X^1_{a,b,c})_{n_1^{(1)} \times n_1^{(2)} \times d_1}$$

to

$$Y^1 = (Y^1_{a,b,c})_{n_1^{(1)} \times n_1^{(2)} \times d_0}$$

between the first and second hidden convolutional layers. We set the d_0 filters with size $1 \times 1 \times d_1$, the stride 1, and define the weights W' and biases B' in this intermediate convolutional layer as

$$W'^{1,c}_{1,1,k} = \begin{cases} p \cdot (-1)^{i+1}, \text{ if } k = (c-1)p + i, \;\; 1 \leq c \leq d_0; \\ 0, \text{ otherwise} \end{cases} \tag{44}$$

and

$$B'^{1,k} = 0 \;\; \forall \;\; 1 \leq k \leq d_0. \tag{45}$$

Then by (2) in the main paper,

$$Y^1_{a,b,c} = p \sum_{i=1}^{p} (-1)^{i+1} X^1_{a,b,(c-1)p+i} \tag{46}$$

for $1 \leq a \leq n_1^{(1)}$, $1 \leq b \leq n_1^{(2)}$, $1 \leq c \leq d_0$. Therefore, (46) determines an affine function

$$Y^1 = \Phi_1'(X^1) \tag{47}$$

from $\mathbb{R}^{n_1^{(1)} \times n_1^{(2)} \times d_1}$ to $\mathbb{R}^{n_1^{(1)} \times n_1^{(2)} \times d_0}$. Therefore, we obtain

$$\begin{aligned} Y^1_{a,b,c} &= p \sum_{i=1}^{p} (-1)^{i+1} X^1_{a,b,(c-1)p+i} \\ &= p \sum_{i=1}^{p} (-1)^{i+1} \max\{0, Z^1_{a,b,(c-1)p+i}\} \\ &= \sum_{i=1}^{p} (-1)^{i+1} \psi_i(p X^0_{1+(a-1)s_1, 1+(b-1)s_1, c}) \\ &= \phi_p(X^0_{1+(a-1)s_1, 1+(b-1)s_1, c}). \end{aligned} \tag{48}$$

The third equality holds due to Eqs. (38) and (41). By the previous discussion on properties of the function $\phi_j(x)$, the following map $\Psi_1 = \Phi_1' \circ \Phi_1$ determined by Eq. (48)

$$\begin{array}{ccccc} \Psi_1 : \mathbb{R}^{n_0^{(1)} \times n_0^{(2)} \times d_0} & \xrightarrow{\Phi_1} & \mathbb{R}^{n_1^{(1)} \times n_1^{(2)} \times d_1} & \xrightarrow{\Phi_1'} & \mathbb{R}^{n_1^{(1)} \times n_1^{(2)} \times d_0} \\ X^0 & \mapsto & X^1 & \mapsto & Y^1 \end{array}$$

sends $\lfloor \frac{d_1}{d_0} \rfloor^{n_1^{(1)} \times n_1^{(2)} \times d_0} = p^{n_1^{(1)} \times n_1^{(2)} \times d_0}$ distinct hypercubes

$$\left\{ [0, \frac{1}{2p}], [\frac{1}{2p}, \frac{2}{2p}], \cdots, [\frac{p-1}{2p}, \frac{p}{2p}] \right\}^{n_0^{(1)} \times n_0^{(2)} \times d_0}$$

in $[0, \frac{1}{2}]^{n_0^{(1)} \times n_0^{(2)} \times d_0}$ onto the same hypercube $[0, \frac{1}{2}]^{n_1^{(1)} \times n_1^{(2)} \times d_0}$ of the intermediate layer $Y^1 \in \mathbb{R}^{n_1^{(1)} \times n_1^{(2)} \times d_0}$ (this map is affine and bijective when restricted to each of the $\left\lfloor \frac{d_1}{d_0} \right\rfloor^{n_1^{(1)} \times n_1^{(2)} \times d_0}$ distinct hypercubes). Similarly (keep d_0 unchanged, and replace $n_0^{(1)}, n_0^{(2)}, n_1^{(1)}, n_1^{(2)}, d_1$ in Ψ_1 by $n_{l-1}^{(1)}, n_{l-1}^{(2)}, n_l^{(1)}, n_l^{(2)}, d_l$), we can define Φ_l, Φ_l', Ψ_l and Y^l for $2 \leq l \leq L-1$ such that the map

$$\begin{array}{ccccc} \Psi_l : \mathbb{R}^{n_{l-1}^{(1)} \times n_{l-1}^{(2)} \times d_0} & \xrightarrow{\Phi_l} & \mathbb{R}^{n_l^{(1)} \times n_l^{(2)} \times d_l} & \xrightarrow{\Phi_l'} & \mathbb{R}^{n_l^{(1)} \times n_l^{(2)} \times d_0} \\ Y^{l-1} & \mapsto & X^{l-1} & \mapsto & Y^l \end{array}$$

sends $\lfloor \frac{d_l}{d_0} \rfloor^{n_l^{(1)} \times n_l^{(2)} \times d_0}$ distinct hypercubes

$$\left\{[0, \frac{1}{2p}], [\frac{1}{2p}, \frac{2}{2p}], \cdots, [\frac{p-1}{2p}, \frac{p}{2p}]\right\}^{n_{l-1}^{(1)} \times n_{l-1}^{(2)} \times d_0}$$

in $[0, \frac{1}{2}]^{n_{l-1}^{(1)} \times n_{l-1}^{(2)} \times d_0}$ onto the hypercube $[0, \frac{1}{2}]^{n_l^{(1)} \times n_l^{(2)} \times d_0}$ of the intermediate layer $Y^l \in \mathbb{R}^{n_l^{(1)} \times n_l^{(2)} \times d_0}$. Therefore,

$$\begin{aligned} \Psi_{L-1} \circ \Psi_{L-2} \circ \cdots \circ \Psi_2 \circ \Psi_1 : \mathbb{R}^{n_0^{(1)} \times n_0^{(2)} \times d_0} &\to \mathbb{R}^{n_{L-1}^{(1)} \times n_{L-1}^{(2)} \times d_0} \\ X^0 &\mapsto Y^{L-1} \end{aligned}$$

sends $\prod_{l=1}^{L-1} \left\lfloor \frac{d_l}{d_0} \right\rfloor^{n_l^{(1)} \times n_l^{(2)} \times d_0}$ distinct hypercubes in $[0, \frac{1}{2}]^{n_0^{(1)} \times n_0^{(2)} \times d_0}$ onto the same hypercube $[0, \frac{1}{2}]^{n_{L-1}^{(1)} \times n_{L-1}^{(2)} \times d_0}$ of the intermediate layer. Note that $\Phi_l \circ \Phi'_{l-1}$ is the convolutional layer between X^{l-1} and X^l which has d_l filter with size $f_l^{(1)} \times f_l^{(2)} \times d_{l-1}$ and stride s_l due to Theorem 4. Finally, by Theorem 2, a one-layer ReLU CNN with input dimension $n_{L-1}^{(1)} \times n_{L-1}^{(2)} \times d_0$ and output dimension $n_L^{(1)} \times n_L^{(2)} \times d_L$ can divide the hypercube $[0, \frac{1}{2}]^{n_{L-1}^{(1)} \times n_{L-1}^{(2)} \times d_0}$ into $R_{\mathcal{N}'}$ regions. Put the network from X^0 to Y^{L-1} and Y^{L-1} to X^L together, we prove the lower bound claim.

(ii) We will prove this claim by induction on L. When $L = 1$, by Theorem 2 the claim is true. Now suppose that $L \geq 2$ and the claim is true for $L-1$. Let $\mathcal{N}^*$ be the CNN obtained from $\mathcal{N}$ by deleting the L-th hidden layer (i.e., $\mathcal{N}^*$ consists of the first to the $L-1$-th layer of $\mathcal{N}$). Then by induction hypothesis, we have

$$R_{\mathcal{N}^*} \leq R_{\mathcal{N}''} \prod_{l=2}^{L-1} \sum_{i=0}^{n_0^{(1)} n_0^{(2)} d_0} \binom{n_l^{(1)} n_l^{(2)} d_l}{i}.$$

Now we consider the L-th layer. Suppose that the CNN $\mathcal{N}^*$ with parameters θ partitions the input space into m distinct linear regions $\mathcal{R}_i$ $(1 \leq i \leq m)$. Since each linear region $\mathcal{R}_i$ corresponds to a certain activation pattern, the function $\mathcal{F}_{\mathcal{N}',\theta}$ becomes an affine function when restricted to $\mathcal{R}_i$. Therefore, after adding the L-th layer to $\mathcal{N}^*$, when restricted to $\mathcal{R}_i$, the function $\mathcal{F}_{\mathcal{N},\theta} \mid_{\mathcal{R}_i}$ can be realised as a one-layer NN with $n_0^{(1)} n_0^{(2)} d_0$ input neurons and $n_l^{(1)} n_l^{(2)} d_l$ hidden neurons. By Proposition 1, $\mathcal{N}$ partitions $\mathcal{R}_i$ into $\sum_{i=0}^{n_0^{(1)} n_0^{(2)} d_0} \binom{n_L^{(1)} n_L^{(2)} d_L}{i}$ distinct linear regions. Finally, we obtain

$$R_{\mathcal{N}} \leq R_{\mathcal{N}^*} \sum_{i=0}^{n_0^{(1)} n_0^{(2)} d_0} \binom{n_L^{(1)} n_L^{(2)} d_L}{i} \leq R_{\mathcal{N}''} \prod_{l=2}^{L} \sum_{i=0}^{n_0^{(1)} n_0^{(2)} d_0} \binom{n_l^{(1)} n_l^{(2)} d_l}{i},$$

which completes the proof. □

4. Calculation of the Number of Parameters for CNNs

Proof of Lemma 4 in the main paper. For the l-th layer, the k-th weight matrix $W^{l,k}$ has $f_l^{(1)} \times f_l^{(2)} \times d_{l-1}$ entries and there are d_l such weight matrices. The bias vector has length d_l. Thus there are $f_l^{(1)} \times f_l^{(2)} \times d_{l-1} \times d_l + d_l$ parameters in the l-th hidden layer. Let l range from 1 to L, the total number of parameters equals $\sum_{l=1}^{L} \left(f_l^{(1)} \times f_l^{(2)} \times d_{l-1} \times d_l + d_l \right)$. □

5. More Examples on the Maximal Number of Linear Regions for One-Layer ReLU CNNs

In this section, we list more examples on maximal number of linear regions for one-layer ReLU CNNs from Tables 1 to 5, which is calculated according to Theorem 2 in the main paper.

Table 1. The results for the maximal number of linear regions for a one-layer ReLU CNN with input dimension $2 \times 2 \times 1$, d_1 filters with dimension $1 \times 2 \times 1$, stride $s = 1$, and hidden layer dimension $2 \times 1 \times d_1$.

	$d_1 = 1$	$d_1 = 2$	$d_1 = 3$	$d_1 = 4$	$d_1 = 5$	$d_1 = 6$	$d_1 = 7$	$d_1 = 8$
$R_{\mathcal{N}}$ by Theorem 2	4	16	49	121	256	484	841	1369
Upper bounds by Theorem 1	4	16	57	163	386	794	1471	2517
Naive upper bounds	4	16	64	256	1024	4096	16384	65536

Table 2. The results for the maximal number of linear regions for a one-layer ReLU CNN with input dimension $1 \times 4 \times 1$, d_1 filters with dimension $1 \times 2 \times 1$, stride $s = 1$, and hidden layer dimension $1 \times 3 \times d_1$.

	$d_1 = 1$	$d_1 = 2$	$d_1 = 3$	$d_1 = 4$	$d_1 = 5$	$d_1 = 6$	$d_1 = 7$	$d_1 = 8$
$R_{\mathcal{N}}$ by Theorem 2	8	55	217	611	1396	2773	4985	8317
Upper bounds by Theorem 1	8	57	256	794	1941	4048	7547	12951
Naive upper bounds	8	64	512	4096	32768	262144	2097152	16777216

Table 3. The results for the maximal number of linear regions for a one-layer ReLU CNN with input dimension $2 \times 3 \times 1$, d_1 filters with dimension $2 \times 2 \times 1$, stride $s = 1$, and hidden layer dimension $2 \times 1 \times d_1$.

	$d_1 = 1$	$d_1 = 2$	$d_1 = 3$	$d_1 = 4$	$d_1 = 5$	$d_1 = 6$	$d_1 = 7$	$d_1 = 8$
$R_{\mathcal{N}}$ by Theorem 2	4	16	64	247	836	2424	6126	13829
Upper bounds by Theorem 1	4	16	64	247	848	2510	6476	14893
Naive upper bounds	4	16	64	256	1024	4096	16384	65536

Table 4. The results for the maximal number of linear regions for a one-layer ReLU CNN with input dimension $6 \times 6 \times 1$, d_1 filters with dimension $1 \times 3 \times 1$, stride $s = 2$, and hidden layer dimension $3 \times 2 \times d_1$.

	$d_1 = 1$	$d_1 = 2$	$d_1 = 3$	$d_1 = 4$	$d_1 = 5$	$d_1 = 6$	$d_1 = 7$	$d_1 = 8$
$R_{\mathcal{N}}$ by Theorem 2	64	4096	250047	9129329	191102976	2537716544	23664622311	167557540697
Upper bounds by Theorem 1	64	4096	262144	16777216	1073741824	68719476736	4398045536122	281443698512817
Naive upper bounds	64	4096	262144	16777216	1073741824	68719476736	4398046511104	281474976710656

Table 5. The results for the maximal number of linear regions for a one-layer ReLU CNN with input dimension $3 \times 3 \times 2$, d_1 filters with dimension $2 \times 2 \times 2$, stride $s = 1$, and hidden layer dimension $2 \times 2 \times d_1$.

	$d_1 = 1$	$d_1 = 2$	$d_1 = 3$	$d_1 = 4$	$d_1 = 5$	$d_1 = 6$	$d_1 = 7$	$d_1 = 8$
$R_{\mathcal{N}}$ by Theorem 2	16	256	4096	65536	1048555	16721253	256376253	3459170397
Upper bounds by Theorem 1	16	256	4096	65536	1048555	16721761	256737233	3485182163
Naive upper bounds	16	256	4096	65536	1048576	16777216	268435456	4294967296

References

Lojasiewicz, S. Triangulation of semi-analytic sets. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 18 (4):449–474, 1964.

Shafarevich, I. R. and Remizov, A. O. *Linear algebra and geometry*. Springer Science & Business Media, 2012.

Stanley, R. P. An introduction to hyperplane arrangements. In *Lecture Notes, IAS/Park City Mathematics Institute*, 2004.

Zaslavsky, T. *Facing up to arrangements : face-count formulas for partitions of space by hyperplanes*. Number 154 in Memoirs of the American Mathematical Society. American Mathematical Society, 1975.