A. Existence of Covering Templates

In this paper we analyze the expressiveness of networks, *i.e.*the functions they can realize, through the notion of *grid tensors*. Recall from sec. 4 that given *templates* $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$, the grid tensor of a score function $h_y : (\mathbb{R}^s)^N \to \mathbb{R}$ realized by some network, is defined to be a tensor of order N and dimension M in each mode, denoted $\mathcal{A}(h_y)$, and given by eq. 3. In particular, it is a tensor holding the values of h_y on all instances $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in (\mathbb{R}^s)^N$ whose *patches* $\mathbf{x}_i$ are taken from the set of templates $\{\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)}\}$ (recurrence allowed). Some of the claims in our analysis (sec. 5) assumed that there exist templates for which grid tensors fully define score functions. That is to say, there exist templates such that score function values outside the exponentially large grid $\{X_{d_1 \ldots d_N} := (\mathbf{x}^{(d_1)}, \ldots, \mathbf{x}^{(d_N)}) : d_1 \ldots d_N \in [M]\}$ are irrelevant for classification. Templates meeting this property were referred to as *covering* (see sec. 5.1). In this appendix we address the existence of covering templates.

If we allow M to grow arbitrarily large then obviously covering templates can be found. However, since in our construction M is tied to the number of channels in the first (representation) layer of a network (see fig. 1), such a trivial observation does not suffice, and in fact we would like to show that covering templates exist for values of M that correspond to practical network architectures, *i.e.* $M \in \Omega(100)$. For such an argument to hold, assumptions must be made on the distribution of input data. Given that ConvNets are used primarily for processing natural images, we assume here that data is governed by their statistics. Specifically, we assume that an instance $X = (\mathbf{x}_1, \ldots, \mathbf{x}_N) \in (\mathbb{R}^s)^N$ corresponds to a natural image, represented through N image patches around its pixels: $\mathbf{x}_1 \ldots \mathbf{x}_N \in \mathbb{R}^s$.

If the dimension of image patches is small then it seems reasonable to believe that relatively few templates can indeed cover the possible appearances of a patch. For example, in the extreme case where each patch is simply a gray-scale pixel ($s = 1$), having $M = 256$ templates may provide the standard 8-bit resolution, leading grid tensors to fully define score functions by accounting for all possible images. However, since in our construction input patches correspond to the receptive field in the first layer of a ConvNet (see fig. 1), we would like to establish an argument for image patch sizes that more closely correlate to typical receptive fields, *e.g.* $5{\times}5$. For this we rely on various studies (*e.g.* (Zoran and Weiss, 2012)) characterizing the statistics of natural images, which have shown that for large ensembles of images, randomly cropped patches of size up to $16{\times}16$ may be relatively well captured by Gaussian Mixture Models with as few as 64 components. This complies with the common belief that there is a moderate number of appearances taken by the vast majority of local image patches (edges, Gabor filters *etc.*). That is to say, it complies with our assumption that covering templates exist with a moderate value of M. We refer the reader to (Cohen et al., 2016b) for a more formal argument on this line.

B. Universality of Fully-Connected Networks

In claim 6 we considered a network obtained by expanding the conv receptive field in the shallow ConvNet, and have shown that it is not universal with ReLU activation and average pooling. As stated thereafter, this result does not contradict the known universality of shallow (single hidden layer) fully-connected neural networks. Resolving the tension is the purpose of this appendix. A shallow fully-connected network corresponds to the shallow ConvNet (fig. 2) with conv receptive field expanded to cover the entire spatial extent, thereby effectively removing the pooling operator (assuming the latter realizes the identity on singletons). In claim 12 below we show that such a network, when equipped with ReLU activation, is universal. In claim 6 on the other hand we assumed that the conv receptive field covers less than half the spatial extent ($w{\cdot}h < N/2 + 1 - \log_M N$), and have shown that with ReLU activation and average pooling, this leads to non-universality. Loosely speaking, our findings imply that for networks with ReLU activation, which are known to be universal when fully-connected, introducing locality disrupts universality with average pooling (while maintaining it with max pooling – claim 4).

Claim 12. *Assume that there exist covering templates $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)}$, and corresponding representation functions $f_{\theta_1} \ldots f_{\theta_M}$ leading to a matrix F (eq. 4) that has non-recurring rows and a constant non-zero column* [6]*. Consider the fully-connected network obtained by expanding the conv receptive field in the shallow ConvNet to cover the entire spatial extent. Such a network, when equipped with ReLU activation, is universal.*

Proof. See app. E.14. □

C. Depth Efficiency with Approximation

In sec. 5.3.1 we stated that the results in our analysis establishing depth efficiency (claims 7, 8, 10 and the analogous ones in app. D), which are currently framed in the context of exact realization, may readily be strengthened to account for arbitrarily-well approximation as well. An explanation for this follows.

When proving that a grid tensor generated by a shallow ConvNet beneath a certain size cannot be equal to a grid tensor generated by a deep ConvNet, we always rely on matricization rank. Namely, we arrange the grid tensors as matrices, and derive constants $R, r \in \mathbb{N}$, $R > r$, such that the matrix corresponding to the deep ConvNet has rank at least R, while that corresponding to the shallow ConvNet has rank at most r. While used in our proofs solely to show that the matrices are different, this actually entails information regarding the distance between them. Namely, if we denote the singular values of the matrix corresponding to the deep ConvNet by $\sigma_1 \geq \sigma_2 \geq \ldots \geq 0$, the squared Euclidean (Frobenius) distance between the matrices is at least $\sigma_{r+1}^2 + \cdots + \sigma_R^2$. Since the matrices are merely rearrangements of the grid tensors, we have a lower bound on the distance between the shallow ConvNet's grid tensor and the target grid tensor generated by the deep ConvNet, so in particular arbitrarily-well approximation is not possible.

D. Shared Coefficients for Convolution

In this appendix we provide our analysis of the shared setting, briefly summarized in sec. 5.4. The analysis follows the same line as that of the unshared setting given in sec. 5.2 and 5.3. For brevity, we assume the reader is familiar with the latter, and do not repeat discussions given there.

[6] The assumption that such representation functions exist differs from our usual non-degeneracy assumption. However, through a slight modification of the proof of claim 1, one can show that standard neurons meet non only non-degeneracy, but also the assumption made here.

As described in sec. 4, the shared setting refers to the case where the 1×1 conv filters in our networks are spatially invariant, giving rise to standard convolutions (as opposed to the more general locally-connected operators). Specifically, the shallow ConvNet (fig. 2) would have a single weight vector $\mathbf{a}^z$ for every hidden channel z, as opposed to the unshared setting where it has a weight vector $\mathbf{a}^{z,i}$ for every location i in every hidden channel z. Grid tensors produced by the shallow ConvNet in the shared setting are given by what we call the *shared generalized CP decomposition*:

$$\mathcal{A}\left(h_y^S\right) = \sum_{z=1}^{Z} a_z^y \cdot \underbrace{(F\mathbf{a}^z) \otimes_g \cdots \otimes_g (F\mathbf{a}^z)}_{N \text{ times}} \tag{8}$$

As for the deep ConvNet (fig. 1 with size-2 pooling windows and $L = \log_2 N$ hidden layers), in the shared setting, instead of having a weight vector $\mathbf{a}^{l,j,\gamma}$ for every hidden layer l, channel γ and location j, there is a single weight vector $\mathbf{a}^{l,\gamma}$ for all locations of channel γ in hidden layer l. Produced grid tensors are then given by the *shared generalized HT decomposition*:

$$\begin{aligned}
\phi^{1,\gamma} &= \sum_{\alpha=1}^{r_0} a_\alpha^{1,\gamma} (F\mathbf{a}^{0,\alpha}) \otimes_g (F\mathbf{a}^{0,\alpha}) \\
&\cdots \\
\phi^{l,\gamma} &= \sum_{\alpha=1}^{r_{l-1}} a_\alpha^{l,\gamma} \underbrace{\phi^{l-1,\alpha}}_{\text{order } 2^{l-1}} \otimes_g \underbrace{\phi^{l-1,\alpha}}_{\text{order } 2^{l-1}} \\
&\cdots \\
\phi^{L-1,\gamma} &= \sum_{\alpha=1}^{r_{L-2}} a_\alpha^{L-1,\gamma} \underbrace{\phi^{L-2,\alpha}}_{\text{order } \frac{N}{4}} \otimes_g \underbrace{\phi^{L-2,\alpha}}_{\text{order } \frac{N}{4}} \\
\mathcal{A}\left(h_y^D\right) &= \sum_{\alpha=1}^{r_{L-1}} a_\alpha^{L,y} \underbrace{\phi^{L-1,\alpha}}_{\text{order } \frac{N}{2}} \otimes_g \underbrace{\phi^{L-1,\alpha}}_{\text{order } \frac{N}{2}} \qquad (9)
\end{aligned}$$

We now turn to analyze universality and depth efficiency in the shared setting.

D.1. Universality

In the unshared setting we saw (sec. 5.2) that linear activation with product pooling and ReLU activation with max pooling both lead to universality, whereas ReLU activation with average pooling does not. We will now see that in the shared setting, no matter how the activation and pooling operators are chosen, universality is never met.

A shallow ConvNet with shared weights produces grid tensors through the shared generalized CP decomposition (eq. 8). A tensor $\mathcal{A}$ generated by this decomposition is necessarily *symmetric*, *i.e.*for any permutation $\delta : [N] \to [N]$ and indexes $d_1 \ldots d_N$ it meets: $\mathcal{A}_{d_1 \ldots d_N} = \mathcal{A}_{\delta(d_1) \ldots \delta(d_N)}$. Obviously not all tensors share this property, so indeed a shallow ConvNet with weight sharing is not universal. A deep ConvNet with weight sharing produces grid tensors through the shared generalized HT decomposition (eq. 9). For this decomposition, a generated tensor $\mathcal{A}$ is invariant to replacing the first and second halves of its modes, *i.e.*for any indexes $d_1 \ldots d_N$ it meets: $\mathcal{A}_{d_1,\ldots,d_N} = \mathcal{A}_{d_{N/2+1},\ldots,d_N,d_1,\ldots,d_{N/2}}$. Although this property is much less stringent than symmetry, it is still not met by most tensors, and so a deep ConvNet with weight sharing is not universal either.

D.2. Depth Efficiency

Depth efficiency deals with the computational complexity of replicating a deep network's function using a shallow network. In order for this question to be applicable, we require that the shallow network be a universal machine. If this is not the case, then it is generally likely that the deep network's function simply lies outside the reach of the shallow network, and we do not obtain a quantitative insight into the true power of depth. Since our shallow ConvNets are not universal with shared weights (app. D.1), we evaluate depth efficiency of deep ConvNets with shared weights against shallow ConvNets with *unshared* weights. Specifically, we do this for the activation-pooling choices leading shallow ConvNets with unshared weights to be universal: linear activation with product pooling, and ReLU activation with max pooling (see sec. 5.2).

For linear activation with product pooling, the following claim, which is essentially a derivative of theorem 1 in (Cohen et al., 2016b), tells us that in the shared setting, as in the unshared setting, depth efficiency holds completely:

Claim 13 (shared analogy of claim 7)**.** *Let $f_{\theta_1} \ldots f_{\theta_M}$ be any set of linearly independent representation functions for a deep ConvNet with linear activation, product pooling and weight sharing. Suppose we randomize the weights of the network by some continuous distribution. Then, with probability 1, we obtain score functions that cannot be realized by a shallow ConvNet with linear activation and product pooling (*not *limited by weight sharing), if the number of hidden channels in the latter (Z) is less than* $\min\{r_0, M\}^{N/2}$.

Proof. See app. E.15. □

Heading on to ReLU activation and max pooling, we will show that here too, the situation in the shared setting is the same as in the unshared setting. Specifically, depth efficiency holds, but not completely. We prove this via two claims, analogous to claims 8 and 9 in sec. 5.3:

Claim 14 (shared analogy of claim 8)**.** *There exist weight settings for a deep ConvNet with ReLU activation, max pooling and weight sharing, giving rise to score functions that cannot be realized by a shallow ConvNet with ReLU activation and max pooling (*not *limited by weight sharing), if the number of hidden channels in the latter (Z) is less than* $\min\{r_0, M\}^{N/2} \cdot \frac{2}{M \cdot N}$.

Proof. See app. E.16. □

Claim 15 (shared analogy of claim 9)**.** *Suppose we randomize the weights of a deep ConvNet with ReLU activation, max pooling and weight sharing by some continuous distribution with non-vanishing continuous probability density function. Then, assuming covering templates exist, with positive probability, we obtain score functions that can be realized by a shallow ConvNet with ReLU activation and max pooling having only a single hidden channel ($Z = 1$).*

Proof. See app. E.17. □

To recapitulate this appendix, we have shown that introducing weight sharing into the 1×1 conv operators of our networks, thereby limiting the general locally-connected linear mappings to be standard convolutions, disrupts universality, but leaves depth efficiency intact – it remains to hold completely under linear activation with product pooling, and incompletely under ReLU activation with max pooling.

E. Deferred Proofs

In this appendix we present the proofs omitted from the text. Beforehand, as preparatory steps, we briefly lay out preliminary material required in order to follow our proofs (app. E.1), and then turn to discuss the important concept of matricization (app. E.2), which facilitates the use of matrix theory for analyzing generalized tensor decompositions.

E.1. Preliminaries

For evaluating the completeness of depth efficiency, and for other purposes as well, we are often interested in the "volume" of sets in a Euclidean space, or more formally, in their Lebesgue measure. While an introduction to Lebesgue measure theory is beyond the scope of this paper (the interested reader is referred to (Jones, 2001)), we restate here several concepts and results our proofs will rely upon. A zero measure set can intuitively be thought of as having zero volume. A union of countably many zero measure sets is itself a zero measure set. If we randomize a point in space by some continuous distribution, the probability of hitting a zero measure set is always zero. A useful fact (proven in (Caron and Traynor, 2005) for example) is that the zero set of a polynomial, *i.e.*the set of points on which a polynomial vanishes, is either the entire space (when the polynomial in question is the zero polynomial), or it must have measure zero. An open set always has positive measure, and when a point in space is drawn by a continuous distribution with non-vanishing continuous probability density function, the probability of hitting such a set is positive.

Apart from measure theory, we will also be using tools from the field of tensor analysis. Here too, a full introduction to the topic is beyond our scope (we refer the interested reader to (Hackbusch, 2012)), and we only list some concepts and results that will be used. First, a fact that relates to abstract tensor products over function spaces is the following. If $f_{\theta_1} \ldots f_{\theta_M} : \mathbb{R}^s \to \mathbb{R}$ are linearly independent functions, then the product functions $\{(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}) \mapsto \prod_{i=1}^{M} f_{\theta_{d_i}}(\mathbf{x}^{(i)})\}_{d_1 \ldots d_M \in [M]}$ from $(\mathbb{R}^s)^M$ to $\mathbb{R}$ are linearly independent as well. Back to tensors as we have defined them (multi-dimensional arrays), a very important concept is that of *rank*, which for order-2 tensors reduces to the standard notion of matrix rank. A tensor is said to have rank 1 if it may be written as a tensor product between non-zero vectors ($\mathcal{A} = \mathbf{v}^1 \otimes \cdots \otimes \mathbf{v}^N$). The rank of a general tensor is defined to be the minimal number of rank-1 tensors that may be summed up to produce it. A useful fact is that the rank of an order-N tensor with dimension M_i in each mode $i \in [N]$, is no greater than $\prod_i M_i / \max_i M_i$. On the other hand, all such tensors, besides a zero measure set, have rank equal to at least $\min\{\prod_{i\ even} M_i, \prod_{i\ odd} M_i\}$. As in the special case of matrices, the rank is sub-additive, *i.e.* $rank(\mathcal{A} + \mathcal{B}) \leq rank(\mathcal{A}) + rank(\mathcal{B})$ for any tensors $\mathcal{A}, \mathcal{B}$ of matching dimensions. The rank is sub-multiplicative w.r.t. the tensor product, *i.e.* $rank(\mathcal{A} \otimes \mathcal{B}) \leq rank(\mathcal{A}) \cdot rank(\mathcal{B})$ for any tensors $\mathcal{A}, \mathcal{B}$. Finally, we use the fact that permuting the modes of a tensor does not alter its rank.

E.2. Matricization

When analyzing grid tensors, we will often consider their arrangement as matrices. The *matricization* of a tensor $\mathcal{A}$, denoted $[\mathcal{A}]$, is its arrangement as a matrix with rows corresponding to odd modes and columns corresponding to even modes. Specifically, if $\mathcal{A} \in \mathbb{R}^{M_1 \times \cdots \times M_N}$, and assuming for simplicity that the order N is even, the matricization $[\mathcal{A}] \in \mathbb{R}^{(M_1 \cdot M_3 \cdot \ldots \cdot M_{N-1}) \times (M_2 \cdot M_4 \cdot \ldots \cdot M_N)}$ holds $\mathcal{A}_{d_1, \ldots, d_N}$ in row index $1 + \sum_{i=1}^{N/2} (d_{2i-1} - 1) \prod_{j=i+1}^{N/2} M_{2j-1}$ and column index $1 + \sum_{i=1}^{N/2} (d_{2i} - 1) \prod_{j=i+1}^{N/2} M_{2j}$.

The matrix analogy of the tensor product $\otimes$ (eq. 1) is called the *Kronecker product*, and is denoted by $\odot$. For $A \in \mathbb{R}^{M_1 \times M_2}$ and $B \in \mathbb{R}^{N_1 \times N_2}$, $A \odot B$ is the matrix in $\mathbb{R}^{M_1 N_1 \times M_2 N_2}$ holding $A_{ij} B_{kl}$ in row index $(i-1)N_1 + k$ and column index $(j-1)N_2 + l$. The relation $[\mathcal{A} \otimes \mathcal{B}] = [\mathcal{A}] \odot [\mathcal{B}]$, where $\mathcal{A}$ and $\mathcal{B}$ are arbitrary tensors of even order, implies that the tensor and Kronecker products are indeed analogous, *i.e.*they represent the same operation under tensor and matrix viewpoints, respectively. We generalize the Kronecker product analogously to our generalization of the tensor product (eq. 2). For an associative and commutative binary operator $g(\cdot, \cdot)$, the *generalized Kronecker product* $\odot_g$, is an operator that intakes matrices $A \in \mathbb{R}^{M_1 \times M_2}$ and $B \in \mathbb{R}^{N_1 \times N_2}$, and returns a matrix $A \odot_g B \in \mathbb{R}^{M_1 N_1 \times M_2 N_2}$ holding $g(A_{ij}, B_{kl})$ in row index $(i-1)N_1 + k$ and column index $(j-1)N_2 + l$. The relation between the tensor and Kronecker products holds for their generalized versions as well, *i.e.*$[\mathcal{A} \otimes_g \mathcal{B}] = [\mathcal{A}] \odot_g [\mathcal{B}]$ for arbitrary tensors $\mathcal{A}, \mathcal{B}$ of even order.

Equipped with the matricization operator $[\cdot]$ and the generalized Kronecker product $\odot_g$, we are now in a position to translate the generalized HT decomposition (eq. 7) to an expression for the matricization of a grid tensor generated by the deep ConvNet:

$$
\begin{aligned}
\phi^{1,j,\gamma} \quad &= \quad \sum_{\alpha=1}^{r_0} a_\alpha^{1,j,\gamma} (F\mathbf{a}^{0,2j-1,\alpha}) \otimes_g (F\mathbf{a}^{0,2j,\alpha}) \qquad (10) \\
&\cdots \\
\left[\phi^{l,j,\gamma}\right] \quad &= \quad \sum_{\alpha=1}^{r_{l-1}} a_\alpha^{l,j,\gamma} \underbrace{\left[\phi^{l-1,2j-1,\alpha}\right]}_{M^{2^{l-2}}\text{-by-}M^{2^{l-2}}} \odot_g \underbrace{\left[\phi^{l-1,2j,\alpha}\right]}_{M^{2^{l-2}}\text{-by-}M^{2^{l-2}}} \\
&\cdots \\
\left[\phi^{L-1,j,\gamma}\right] \quad &= \quad \sum_{\alpha=1}^{r_{L-2}} a_\alpha^{L-1,j,\gamma} \underbrace{\left[\phi^{L-2,2j-1,\alpha}\right]}_{M^{N/8}\text{-by-}M^{N/8}} \odot_g \underbrace{\left[\phi^{L-2,2j,\alpha}\right]}_{M^{N/8}\text{-by-}M^{N/8}} \\
\left[\mathcal{A}\left(h_y^D\right)\right] \quad &= \quad \sum_{\alpha=1}^{r_{L-1}} a_\alpha^{L,1,y} \underbrace{\left[\phi^{L-1,1,\alpha}\right]}_{M^{N/4}\text{-by-}M^{N/4}} \odot_g \underbrace{\left[\phi^{L-1,2,\alpha}\right]}_{M^{N/4}\text{-by-}M^{N/4}}
\end{aligned}
$$

We refer to this factorization as the *matricized generalized HT decomposition*. Notice that the expression above for $\phi^{1,j,\gamma}$ is the same as in the original generalized HT decomposition, as order-2 tensors need not be matricized.

For the matricization of a grid tensor generated by the shallow ConvNet, we translate the generalized CP decomposition (eq. 6) into the *matricized generalized CP decomposition*:

$$
\begin{aligned}
&\left[\mathcal{A}\left(h_y^S\right)\right] = \qquad (11) \\
&\sum_{z=1}^{Z} a_z^y \cdot \left((F\mathbf{a}^{z,1}) \odot_g (F\mathbf{a}^{z,3}) \odot_g \cdots \odot_g (F\mathbf{a}^{z,N-1}) \right) \odot_g \\
&\qquad \left((F\mathbf{a}^{z,2}) \odot_g (F\mathbf{a}^{z,4}) \odot_g \cdots \odot_g (F\mathbf{a}^{z,N}) \right)^\top
\end{aligned}
$$

The matricized generalized CP and HT decompositions (eq. 11 and 10 respectively) will be used throughout our proofs to establish depth efficiency. This is generally done by providing a lower bound on $rank[\mathcal{A}(h_y^D)]$ – the rank of the deep ConvNet's matricized grid tensor, and an upper bound on $rank[\mathcal{A}(h_y^S)]$ – the rank of the shallow ConvNet's matricized grid tensor. The upper bound on $rank[\mathcal{A}(h_y^S)]$ will be linear in Z, and so requiring $\mathcal{A}(h_y^S) = \mathcal{A}(h_y^D)$, and in particular $rank[\mathcal{A}(h_y^S)] = rank[\mathcal{A}(h_y^D)]$, will give us a lower bound on Z. That is to say, we obtain a lower bound on the number of hidden channels in the shallow ConvNet, that must be met in order for this network to replicate a grid tensor generated by the deep ConvNet.

E.3. Proof of claim 1

We first show that given distinct $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$, there exists a vector $\mathbf{w} \in \mathbb{R}^s$ such that $\mathbf{w}^\top \mathbf{x}^{(i)} \neq \mathbf{w}^\top \mathbf{x}^{(j)}$ for all $1 \leq i < j \leq M$. $\mathbf{w}$ satisfies this condition if it is not perpendicular to any of the finitely many non-zero vectors $\{\mathbf{x}^{(i)} - \mathbf{x}^{(j)} : 1 \leq i < j \leq M\}$. If for every $1 \leq i < j \leq M$ we denote by $P^{(i,j)} \subset \mathbb{R}^s$ the set of points perpendicular to $\mathbf{x}^{(i)} - \mathbf{x}^{(j)}$, we obtain that $\mathbf{w}$ satisfies the desired condition if it does not lie in the union $\bigcup_{1 \leq i < j \leq M} P^{(i,j)}$. Each $P^{(i,j)}$ is the zero set of a non-zero polynomial, and in particular has measure zero. The finite union $\bigcup_{1 \leq i < j \leq M} P^{(i,j)}$ thus has measure zero as well, and accordingly cannot cover the entire space. This implies that $\mathbf{w} \in \mathbb{R}^s \setminus \bigcup_{1 \leq i < j \leq M} P^{(i,j)}$ indeed exists.

Assume without loss of generality $\mathbf{w}^\top \mathbf{x}^{(1)} < \ldots < \mathbf{w}^\top \mathbf{x}^{(M)}$. We may then choose $b_1 \ldots b_M \in \mathbb{R}$ such that $-\mathbf{w}^\top \mathbf{x}^{(M)} < b_M < \ldots < -\mathbf{w}^\top \mathbf{x}^{(1)} < b_1$. For $i, j \in [M]$, $\mathbf{w}^\top \mathbf{x}^{(i)} + b_j$ is positive when $j \leq i$ and negative when $j > i$. Therefore, if $\sigma(\cdot)$ is chosen as the ReLU activation, defining $f_{\theta_j}(\mathbf{x}) = \sigma(\mathbf{w}^\top \mathbf{x} + b_j)$ for every $j \in [M]$ gives rise to a matrix F (eq. 4) that is lower triangular with non-zero values on its diagonal. This proves the desired result for the case of ReLU activation.

Consider now the case of sigmoidal activation, where $\sigma(\cdot)$ is monotonic with $\lim_{z \to -\infty} \sigma(z) = c$ and $\lim_{z \to +\infty} \sigma(z) = C$ for some $c \neq C$ in $\mathbb{R}$. Letting $\mathbf{w} \in \mathbb{R}^s$ and $b_1 \ldots b_M \in \mathbb{R}$ be as above, we introduce a scaling factor $\alpha > 0$, and define $f_{\theta_j}(\mathbf{x}) = \sigma(\alpha \mathbf{w}^\top \mathbf{x} + \alpha b_j)$ for every $j \in [M]$. It is not difficult to see that as $\alpha \to +\infty$, the matrix F tends closer and closer to a matrix holding C on and below its diagonal, and c elsewhere. The latter matrix is non-singular, and in particular has non-zero determinant $d \neq 0$. The determinant of F converges to d as $\alpha \to +\infty$, so for large enough α, F is non-singular. □

E.4. Proof of claim 2

We may view the determinant of F (eq. 4) as a function of $(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)})$:

$$\det F(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}) = \sum_{\delta \in S_M} sign(\delta) \prod_{i=1}^{M} f_{\theta_{\delta(i)}}(\mathbf{x}^{(i)})$$

where S_M stands for the permutation group on $[M]$, and $sign(\delta) \in \{\pm 1\}$ is the sign of the permutation δ. This in particular shows that $\det F(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)})$ is a non-zero linear combination of the product functions $\{(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}) \mapsto \prod_{i=1}^{M} f_{\theta_{d_i}}(\mathbf{x}^{(i)})\}_{d_1 \ldots d_M \in [M]}$. Since these product functions are linearly independent (see app. E.1), $\det F(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)})$ cannot be the zero function. That is to say, there exist $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$ such that $\det F(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(M)}) \neq 0$. □

E.5. Proof of claim 3

Let $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$ be distinct covering templates, and $f_{\theta_1} \ldots f_{\theta_M}$ be representation functions for which F is invertible (non-degeneracy implies that such functions exist). With linear activation and product pooling the generalized CP decomposition (eq. 6) reduces to its standard version, which is known to be able to express any tensor when size is large enough (*e.g.* $Z \geq M^N$ suffices). The shallow ConvNet can thus realize any grid tensor on covering templates, precisely meaning that it is universal. As for the deep ConvNet, setting $r_0 = \cdots = r_{L-1} = Z$ and $a_\alpha^{l,j,\gamma} = \mathbb{1}[\alpha = \gamma]$, where $l \in [L-1]$ and $\mathbb{1}[\cdot]$ is the indicator function, reduces its decomposition (eq. 7) to that of the shallow ConvNet (eq. 6). This implies that all grid tensors realizable by the shallow ConvNet are also realizable by the deep ConvNet. □

E.6. Proof of claim 4

The proof follows the same line as that of claim 3, except we cannot rely on the ability of the standard CP decomposition to realize any tensor of choice. Instead, we need to show that the generalized CP decomposition (eq. 6) with $g(a,b) = \max\{a,b,0\}$ can realize any tensor, so long as Z is large enough. We will show that $Z \geq 2 \cdot M^N$ suffices. For that, it is enough to consider an arbitrary indicator tensor, *i.e.*a tensor holding 1 in some entry and 0 in all other entries, and show that it can be expressed with $Z = 2$.

Let $\mathcal{A}$ be an indicator tensor of order N and dimension M in each mode, its active entry being $(d_1, \ldots, d_N)$. Denote by $\mathbf{1} \in \mathbb{R}^M$ the vector holding 1 in all entries, and for every $i \in [N]$, let $\bar{\mathbf{e}}_{d_i} \in \mathbb{R}^M$ be the vector holding 0 in entry d_i and 1 elsewhere. With the following weight settings, a generalized CP decomposition (eq. 6) with $g(a,b) = \max\{a,b,0\}$ and $Z = 2$ produces $\mathcal{A}$, as required:

- $a_1^y = 1, a_2^y = -1$
- $\mathbf{a}^{1,1} = \cdots = \mathbf{a}^{1,N} = \mathbf{1}$
- $\forall i \in [N] : \mathbf{a}^{2,i} = \bar{\mathbf{e}}_{d_i}$

□

E.7. Proof of claim 5

Let $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$ be any templates of choice, and consider grid tensors produced by the generalized CP and HT decompositions (eq. 6 and 7 respectively) with $g(a,b) = \max\{a,0\} + \max\{b,0\}$ (this corresponds to *sum* pooling and ReLU activation, but as stated in sec. 4, sum and average pooling are equivalent in terms of expressiveness). We will show that such grid tensors, when arranged as matrices, necessarily have low rank. This obviously implies that they cannot take on any value. Moreover, since the set of low rank matrices has zero measure in the space of all matrices (see app. E.1), the set of values that can be taken by the grid tensors has zero measure in the space of tensors with order N and dimension M in each mode.

In accordance with the above, we complete our proof by showing that with $g(a,b) = \max\{a,0\} + \max\{b,0\}$, the matricized generalized CP and HT decompositions (eq. 11 and 10 respectively) give rise to low-rank matrices. For the matricized generalized CP decomposition (eq. 11), corresponding to the shallow ConvNet, we have with $g(a,b)=\max\{a,0\}+\max\{b,0\}$:

$$\left[\mathcal{A}\left(h_y^S\right)\right]=\mathbf{v}\mathbf{1}^\top+\mathbf{1}\mathbf{u}^\top$$

where $\mathbf{1}$ is the vector in $\mathbb{R}^{M^{N/2}}$ holding 1 in all entries, and $\mathbf{v},\mathbf{u}\in\mathbb{R}^{M^{N/2}}$ are defined as follows:

$$\begin{aligned}\mathbf{v} &:= \sum_{z=1}^{Z} a_z^y\cdot\max\left\{(F\mathbf{a}^{z,1})\odot_g\cdots\odot_g(F\mathbf{a}^{z,N-1}),0\right\}\\ \mathbf{u} &:= \sum_{z=1}^{Z} a_z^y\cdot\max\left\{(F\mathbf{a}^{z,2})\odot_g\cdots\odot_g(F\mathbf{a}^{z,N}),0\right\}\end{aligned}$$

Obviously the matrix $\left[\mathcal{A}\left(h_y^S\right)\right]\in\mathbb{R}^{M^{N/2}\times M^{N/2}}$ has rank 2 or less.

Turning to the matricized generalized HT decomposition (eq. 10), which corresponds to the deep ConvNet, we have with $g(a,b)=\max\{a,0\}+\max\{b,0\}$:

$$\left[\mathcal{A}\left(h_y^D\right)\right]=V\odot O+O\odot U$$

where $\odot$ is the standard Kronecker product (see definition in app. E.2), $O\in\mathbb{R}^{M^{N/4}\times M^{N/4}}$ is a matrix holding 1 in all entries, and the matrices $V,U\in\mathbb{R}^{M^{N/4}\times M^{N/4}}$ are given by:

$$\begin{aligned}V &:= \sum_{\alpha=1}^{r_{L-1}} a_\alpha^{L,1,y}\max\left\{\left[\phi^{L-1,1,\alpha}\right],0\right\}\\ U &:= \sum_{\alpha=1}^{r_{L-1}} a_\alpha^{L,1,y}\max\left\{\left[\phi^{L-1,2,\alpha}\right],0\right\}\end{aligned}$$

The rank of O is obviously 1, and since the Kronecker product multiplies ranks, *i.e.*$rank(A\odot B)=rank(A){\cdot}rank(B)$ for any matrices A and B, we have that the rank of $\left[\mathcal{A}\left(h_y^D\right)\right]\in\mathbb{R}^{M^{N/2}\times M^{N/2}}$ is at most $2{\cdot}M^{N/4}$. In particular, $\left[\mathcal{A}\left(h_y^D\right)\right]$ cannot have full rank. □

E.8. Proof of claim 6

Compare the original shallow ConvNet (fig. 2) to the shallow ConvNet with expanded receptive field that we consider in this claim, illustrated in fig. 3. The original shallow ConvNet has 1×1 receptive field, with conv entry in location $i\in[N]$ and channel $z\in[Z]$ assigned through a cross-channel linear combination of the representation entries in the same location, the combination weights being $\mathbf{a}^{z,i}\in\mathbb{R}^M$. In the shallow ConvNet with receptive field expanded to $w{\times}h$, linear combinations span multiple locations. In particular, conv entry in location i and channel z is now assigned through a linear combination of the representation entries at all channels that lie inside a spatial window revolving around i. We denote by $\{\rho(j;i)\}_{j\in[w\cdot h]}$ the locations comprised by this window. More specifically, $\rho(j;i)$ is the j'th location in the window, and the linear weights that correspond to it are held in the j'th column of the weight matrix $A^{z,i}\in\mathbb{R}^{M\times w\cdot h}$. We assume for simplicity that conv windows stepping out of bounds encounter zero padding [7], and adhere to the convention under which indexing the row of a matrix with $d_{\rho(j;i)}$ produces zero when location j of window i steps out of bounds.

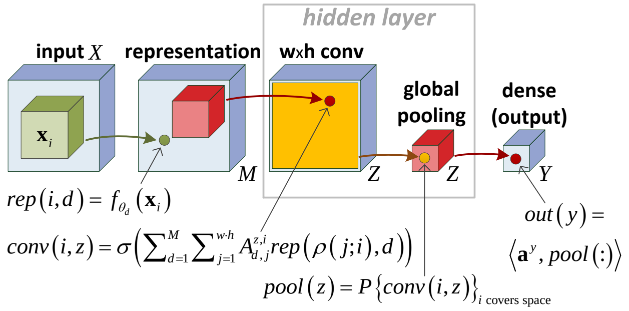
hidden layer
input X
representation
w×h conv
global pooling
dense (output)
$\mathbf{x}_i$
M
Z
Z
Y
$rep(i,d)=f_{\theta_d}(\mathbf{x}_i)$
$conv(i,z)=\sigma\left(\sum_{d=1}^{M}\sum_{j=1}^{w\cdot h}A_{d,j}^{z,i}rep(\rho(j;i),d)\right)$
$pool(z)=P\{conv(i,z)\}_{i\text{ covers space}}$
$out(y)=$
$\langle\mathbf{a}^y,pool(:)\rangle$

Figure 3. Shallow ConvNet with conv receptive field expanded from 1×1 to $w{\times}h$. The weight vectors $\mathbf{a}^{i,z}\in\mathbb{R}^M$ have been replaced by matrices $A^{i,z}\in\mathbb{R}^{M\times w\cdot h}$, and we denote by $\rho(j;i)$ the spatial location of element j in the $w{\times}h$ window revolving around i. Best viewed in color.

We are interested in the case of ReLU activation ($\sigma(z)=\max\{0,z\}$) and average pooling ($P\{c_j\}=\text{mean}\{c_j\}$). Under this setting, for any selected templates $\mathbf{x}^{(1)}\ldots\mathbf{x}^{(M)}\in\mathbb{R}^s$, the grid tensor of $h_y^{S(w\times h)}$ – network's y'th score function, is given by:

$$\mathcal{A}(h_y^{S(w\times h)})_{d_1,\ldots,d_N}=\sum_{i=1}^{N}\mathcal{B}^i_{d_{\rho(1;i)},\ldots,d_{\rho(w\cdot h;i)}}$$

where for every $i\in[N]$, $\mathcal{B}^i$ is a tensor of order $w{\cdot}h$ and dimension M in each mode, defined by:

$$\mathcal{B}^i_{c_1,\ldots,c_{w\cdot h}}=\sum_{z=1}^{Z}\frac{a_z^y}{N}\max\left\{\sum_{j=1}^{w\cdot h}(FA^{z,i})_{c_j,j},0\right\}$$

Let $\mathcal{O}$ be a tensor of order $N-w{\cdot}h$ and dimension M in each mode, holding 1 in all entries. We may write:

$$\mathcal{A}(h_y^{S(w\times h)})=\sum_{i=1}^{N}p_i(\mathcal{B}^i\otimes\mathcal{O})\qquad(12)$$

where for every $i\in[N]$, $p_i(\cdot)$ is an appropriately chosen operator that permutes the modes of an order-N tensor.

We now make use of some known facts related to tensor rank (see app. E.1), in order to show that eq. 12 is not universal, *i.e.*that there are many tensors which cannot be realized by $\mathcal{A}(h_y^{S(w\times h)})$. Being tensors of order $w{\cdot}h$ and dimension M in each mode, the ranks of $\mathcal{B}^1\ldots\mathcal{B}^N$ are bounded above by $M^{w\cdot h-1}$. Since $\mathcal{O}$ is an all-1 tensor, and since permuting modes does not alter rank, we have: $rank(p_i(\mathcal{B}^i\otimes\mathcal{O}))\leq M^{w\cdot h-1}\;\forall i\in[N]$. Finally, from subadditivity of the rank we get: $rank(\mathcal{A}(h_y^{S(w\times h)}))\leq N{\cdot}M^{w\cdot h-1}$. Now, we know by assumption that $w{\cdot}h<{}^N\!/_2+1-\log_M N$, and this implies: $rank(\mathcal{A}(h_y^{S(w\times h)}))<M^{N/2}$. Since there exist tensors of order N and dimension M in each mode having rank at least $M^{N/2}$ (actually only a negligible set of tensors do not meet this), eq. 12 is indeed not universal. That is to say, the shallow ConvNet with conv receptive field expanded to $w{\times}h$ (fig. 3) cannot realize all grid tensors on the templates $\mathbf{x}^{(1)}\ldots\mathbf{x}^{(M)}$. □

[7] Modifying our proof to account for different padding schemes (such as duplication or no padding at all) is trivial – we choose to work with zero padding merely for notational convenience.

E.9. Proof of claim 7

Let $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$ be templates such that F is invertible (existence follows from claim 2). The deep network generates grid tensors on $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)}$ through the standard HT decomposition (eq. 7 with $g(a,b) = a{\cdot}b$). The proof of theorem 1 in (Cohen et al., 2016b) shows that when arranged as matrices, such tensors have rank at least $\min\{r_0, M\}^{N/2}$ almost always, *i.e.*for all weight ($\mathbf{a}^{l,j,\gamma}$) settings but a set of (Lebesgue) measure zero. On the other hand, the shallow network generates grid tensors on $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)}$ through the standard CP decomposition (eq. 6 with $g(a,b) = a{\cdot}b$), possibly with a different matrix F (representation functions need not be the same). Such tensors, when arranged as matrices, are shown in the proof of theorem 1 in (Cohen et al., 2016b) to have rank at most Z. Therefore, for them to realize the grid tensors generated by the deep network, we almost always must have $Z \geq \min\{r_0, M\}^{N/2}$. □

E.10. Proof of claim 8

The proof traverses along the following path. Letting $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$ be any distinct templates, we show that when arranged as matrices, grid tensors on $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)}$ generated by the shallow network have rank at most $Z {\cdot} \frac{M \cdot N}{2}$. Then, defining $f_{\theta_1} \ldots f_{\theta_M}$ to be representation functions for the deep network giving rise to an invertible F (non-degeneracy implies that such functions exist), we show explicit linear weight ($\mathbf{a}^{l,j,\gamma}$) settings under which the grid tensors on $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)}$ generated by the deep network, when arranged as matrices, have rank at least $\min\{r_0, M\}^{N/2}$.

In light of the above, the proof boils down to showing that with $g(a,b) = \max\{a,b,0\}$:

- The matricized generalized CP decomposition (eq. 11) produces matrices with rank at most $Z {\cdot} \frac{M \cdot N}{2}$.
- For an invertible F, there exists a weight ($\mathbf{a}^{l,j,\gamma}$) setting under which the matricized generalized HT decomposition (eq. 10) produces a matrix with rank at least $\min\{r_0, M\}^{N/2}$.

We begin with the first point, showing that for every $\mathbf{v}_1, \ldots, \mathbf{v}_{N/2} \in \mathbb{R}^M$ and $\mathbf{u}_1, \ldots, \mathbf{u}_{N/2} \in \mathbb{R}^M$:

$$rank\left(\mathbf{v}_1 \odot_g \cdots \odot_g \mathbf{v}_{\frac{N}{2}}\right) \odot_g \left(\mathbf{u}_1 \odot_g \cdots \odot_g \mathbf{u}_{\frac{N}{2}}\right)^\top \leq \frac{M \cdot N}{2} \tag{13}$$

This would imply that every summand in the matricized generalized CP decomposition (eq. 11) has rank at most $\frac{M \cdot N}{2}$, and the desired result readily follows. To prove eq. 13, note that each of the vectors $\bar{\mathbf{v}} := \mathbf{v}_1 \odot_g \cdots \odot_g \mathbf{v}_{\frac{N}{2}}$ and $\bar{\mathbf{u}} := \mathbf{u}_1 \odot_g \cdots \odot_g \mathbf{u}_{\frac{N}{2}}$ are of dimension $M^{N/2}$, but have only up to $\frac{M \cdot N}{2}$ unique values. Let $\delta_{\mathbf{v}}, \delta_{\mathbf{u}} : [M^{N/2}] \to [M^{N/2}]$ be permutations that arrange the entries of $\bar{\mathbf{v}}$ and $\bar{\mathbf{u}}$ in descending order. Permuting the rows of the matrix $\bar{\mathbf{v}} \odot_g \bar{\mathbf{u}}^\top$ via $\delta_{\mathbf{v}}$, and the columns via $\delta_{\mathbf{u}}$, obviously does not change its rank. On the other hand, we get a $M^{N/2} {\times} M^{N/2}$ matrix with a $\frac{M \cdot N}{2} \times \frac{M \cdot N}{2}$ block structure, each block being constant (*i.e.*all entries of a block hold the same value). This implies that the rank of $\bar{\mathbf{v}} \odot_g \bar{\mathbf{u}}^\top$ is at most $\frac{M \cdot N}{2}$, which is what we set out to prove.

Moving on to the matricized generalized HT decomposition (eq. 10), for an invertible F we define the following weight setting ($\mathbf{0}$ and $\mathbf{1}$ here denote the all-0 and all-1 vectors, respectively):

- $\mathbf{a}^{0,j,\gamma} = \begin{cases} F^{-1}\bar{\mathbf{e}}_\gamma & ,\gamma{\leq}M \\ \mathbf{0} & ,\gamma > M \end{cases}$, where $\bar{\mathbf{e}}_\gamma \in \mathbb{R}^M$ is defined to be the vector holding 0 in entry γ and 1 in all other entries.
- $\mathbf{a}^{l,j,\gamma} = \begin{cases} \mathbf{1} & ,\gamma = 1 \,,\; l \in [L-1] \\ \mathbf{0} & ,\gamma > 1 \,,\; l \in [L-1] \end{cases}$
- $\mathbf{a}^{L,1,y} = \mathbf{1}$

Under this setting, the produced matrix $\left[\mathcal{A}\left(h_y^D\right)\right]$ holds $\min\{r_0, M\}$ everywhere besides $\min\{r_0, M\}^{N/2}$ entries on its diagonal, where it holds $\min\{r_0, M\} - 1$. The rank of this matrix is at least $\min\{r_0, M\}^{N/2}$. □

E.11. Proof of claim 9

Let $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$ be covering templates, and $f_{\theta_1} \ldots f_{\theta_M}$ be representation functions for the deep network under which F is invertible (non-degeneracy implies that such functions exist). We will show that there exists a linear weight ($\mathbf{a}^{l,j,\gamma}$) setting for the deep network with which it generates a grid tensor that is realizable by a shallow network with a single hidden channel ($Z = 1$). Moreover, we show that when the representation parameters (θ_d) and linear weights ($\mathbf{a}^{l,j,\gamma}$) are subject to small perturbations, the deep network's grid tensor can still be realized by a shallow network with a single hidden channel. Since templates are covering grid tensors fully define score functions. This, along with the fact that open sets in Lebesgue measure spaces always have positive measure (see app. E.1), imply that there is positive measure to the set of weight configurations leading the deep network to generate score functions realizable by a shallow network with $Z = 1$. Translating the latter statement from measure theoretical to probabilistic terms readily proves the result we seek after.

In light of the above, the proof boils down to the following claim, framed in terms of our generalized tensor decompositions. Fixing $g(a,b) = \max\{a,b,0\}$, per arbitrary invertible F there exists a weight ($\mathbf{a}^{l,j,\gamma}$) setting for the generalized HT decomposition (eq. 7), such that the produced tensor may be realized by the generalized CP decomposition (eq. 6) with $Z = 1$, and this holds even if the weights $\mathbf{a}^{l,j,\gamma}$ and matrix F are subject to small perturbations [8].

We will now show that the following weight setting meets our requirement ($\mathbf{0}$ and $\mathbf{1}$ here denote the all-0 and all-1 vectors, respectively):

- $\mathbf{a}^{0,j,\gamma} = \begin{cases} F^{-1}\mathbf{1} & ,j \text{ odd} \\ \mathbf{0} & ,j \text{ even} \end{cases}$
- $\mathbf{a}^{l,j,\gamma} = \begin{cases} \mathbf{1} & ,j \text{ odd} \;\;,\; l \in [L-1] \\ \mathbf{0} & ,j \text{ even} \,,\; l \in [L-1] \end{cases}$
- $\mathbf{a}^{L,1,y} = \mathbf{1}$

[8] Recall that by assumption representation functions are continuous w.r.t.their parameters ($f_\theta(\mathbf{x})$ is continuous w.r.t.θ), and so small perturbations on representation parameters (θ_d) translate into small perturbations on the matrix F (eq. 4).

Let $\mathcal{E}^F$ be an additive noise matrix applied to F, and $\{\boldsymbol{\epsilon}^{l,j,\gamma}\}_{l,j,\gamma}$ be additive noise vectors applied to $\{\mathbf{a}^{l,j,\gamma}\}_{l,j,\gamma}$. We use the notation $\mathbf{o}(\epsilon)$ to refer to vectors that tend to $\mathbf{0}$ as $\mathcal{E}^F \to 0$ and $\boldsymbol{\epsilon}^{l,j,\gamma} \to \mathbf{0}$, with the dimension of a vector to be understood by context. Plugging in the noisy variables into the generalized HT decomposition (eq. 7), we get for every $j \in [N/2]$ and $\alpha \in [r_0]$:

$$\begin{aligned}
&((F+\mathcal{E}^F)(\mathbf{a}^{0,2j-1,\alpha}+\boldsymbol{\epsilon}^{0,2j-1,\alpha}))\\
&\otimes_g((F+\mathcal{E}^F)(\mathbf{a}^{0,2j,\alpha}+\boldsymbol{\epsilon}^{0,2j,\alpha}))\\
&=((F+\mathcal{E}^F)(F^{-1}\mathbf{1}+\boldsymbol{\epsilon}^{0,2j-1,\alpha}))\\
&\otimes_g((F+\mathcal{E}^F)(\mathbf{0}+\boldsymbol{\epsilon}^{0,2j,\alpha}))\\
&=(\mathbf{1}+\mathbf{o}(\epsilon))\otimes_g \mathbf{o}(\epsilon)
\end{aligned}$$

If the applied noise $(\mathcal{E}^F, \boldsymbol{\epsilon}^{l,j,\gamma})$ is small enough this is equal to $(\mathbf{1}+\mathbf{o}(\epsilon)) \otimes \mathbf{1}$ (recall that $\otimes$ stands for the *standard* tensor product), and we in turn get for every $j \in [N/4]$ and $\gamma \in [r_1]$:

$$\begin{aligned}
&\phi^{1,2j-1,\gamma}\otimes_g\phi^{1,2j,\gamma}\\
&=\left(\textstyle\sum_{\alpha=1}^{r_0} a_\alpha^{1,2j-1,\gamma}(\mathbf{1}+\mathbf{o}(\epsilon))\otimes\mathbf{1}\right)\\
&\otimes_g\left(\textstyle\sum_{\alpha=1}^{r_0} a_\alpha^{1,2j,\gamma}(\mathbf{1}+\mathbf{o}(\epsilon))\otimes\mathbf{1}\right)\\
&=\left(\textstyle\sum_{\alpha=1}^{r_0}(1+\epsilon_\alpha^{1,2j-1,\gamma})(\mathbf{1}+\mathbf{o}(\epsilon))\otimes\mathbf{1}\right)\\
&\otimes_g\left(\textstyle\sum_{\alpha=1}^{r_0}\epsilon_\alpha^{1,2j,\gamma}(\mathbf{1}+\mathbf{o}(\epsilon))\otimes\mathbf{1}\right)\\
&=((r_0\mathbf{1}+\mathbf{o}(\epsilon))\otimes\mathbf{1})\otimes_g(\mathbf{o}(\epsilon)\otimes\mathbf{1})
\end{aligned}$$

With the applied noise $(\mathcal{E}^F, \boldsymbol{\epsilon}^{l,j,\gamma})$ small enough this becomes $(r_0\mathbf{1}+\mathbf{o}(\epsilon)\otimes\mathbf{1}\otimes\mathbf{1}\otimes\mathbf{1}$. Continuing in this fashion over the levels of the decomposition, we get that with small enough noise, for every $l \in [L-1]$, $j \in [N/2^{l+1}]$ and $\gamma \in [r_l]$:

$$\phi^{l,2j-1,\gamma}\otimes_g\phi^{l,2j,\gamma}=\left(\prod\nolimits_{l'=0}^{l-1} r_{l'}\cdot\mathbf{1}+\mathbf{o}(\epsilon)\right)\otimes\left(\otimes_{i=1}^{2^{l+1}-1}\mathbf{1}\right)$$

where $\otimes_{i=1}^{2^{l+1}-1}\mathbf{1}$ stands for the tensor product of the vector $\mathbf{1}$ with itself $2^{l+1}-1$ times. We readily conclude from this that with small enough noise, the tensor produced by the decomposition may be written as follows:

$$\mathcal{A}\left(h_y^D\right)=\left(\prod\nolimits_{l=0}^{L-1} r_l\cdot\mathbf{1}+\mathbf{o}(\epsilon)\right)\otimes\left(\otimes_{i=1}^{N-1}\mathbf{1}\right) \qquad (14)$$

To finish our proof, it remains to show that a tensor as in eq. 14 may be realized by the generalized CP decomposition (eq. 6) with $Z = 1$ (and $g(a,b) = \max\{a,b,0\}$). Indeed, we may assume that the latter's F, which we denote by $\tilde{F}$ to distinguish from the matrix in the generalized HT decomposition (eq. 7), is invertible (non-degeneracy ensures that this may be achieved with proper choice of representation functions for the shallow ConvNet). Setting the weights of the generalized CP decomposition (eq. 6) through:

- $a_1^y = 1$
- $\mathbf{a}^{1,i} = \begin{cases} \tilde{F}^{-1}\left(\prod_{l=0}^{L-1} r_l\cdot\mathbf{1}+\mathbf{o}(\epsilon)\right) & ,i=1 \\ \mathbf{0} & ,i>1 \end{cases}$

leads to $\mathcal{A}\left(h_y^S\right) = \mathcal{A}\left(h_y^D\right)$, as required. □

E.12. Proof of claim 10

The proof here follows readily from those of claims 7 and 8. Namely, in the proof of claim 7 we state that for templates $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$ chosen such that F is invertible (these exist according to claim 2), a grid tensor produced by the deep ConvNet with linear activation and product pooling, when arranged as a matrix, has rank at least $\min\{r_0, M\}^{N/2}$ for all linear weight $(\mathbf{a}^{l,j,\gamma})$ settings but a set of measure zero. That is to say, a matrix produced by the matricized generalized HT decomposition (eq. 10) with $g(a,b) = a{\cdot}b$, has rank at least $\min\{r_0, M\}^{N/2}$ for all weight $(\mathbf{a}^{l,j,\gamma})$ settings but a set of measure zero. On the other hand, we have shown in the proof of claim 8 that a shallow ConvNet with ReLU activation and max pooling generates grid tensors that when arranged as matrices, have rank at most $Z{\cdot}\frac{M\cdot N}{2}$. More specifically, we have shown that the matricized generalized CP decomposition (eq. 11) with $g(a,b) = \max\{a,b,0\}$ produces matrices with rank at most $Z{\cdot}\frac{M\cdot N}{2}$. This implies that under almost all linear weight $(\mathbf{a}^{l,j,\gamma})$ settings for a deep ConvNet with linear activation and product pooling, the generated grid tensor cannot be replicated by a shallow ConvNet with ReLU activation and max pooling if the latter has less than $Z = \min\{r_0, M\}^{N/2}\cdot\frac{2}{M\cdot N}$ hidden channels. □

E.13. Proof of claim 11

The proof here is almost identical to that of claim 9. The only difference is where we show that a tensor as in eq. 14 may be realized by the generalized CP decomposition (eq. 6) with $Z = 1$. In the proof of claim 9 the underlying operation of the decomposition was $g(a,b) = \max\{a,b,0\}$ (corresponding to ReLU activation and max pooling), whereas here it is $g(a,b) = a{\cdot}b$ (corresponding to linear activation and product pooling). To account for this difference, we again assume that $\tilde{F}$ – the matrix F of the generalized CP decomposition, is invertible (non-degeneracy ensures that this may be achieved with proper choice of representation functions for the shallow ConvNet), and modify the decomposition's weight setting as follows:

- $a_1^y = 1$
- $\mathbf{a}^{1,i} = \begin{cases} \tilde{F}^{-1}\left(\prod_{l=0}^{L-1} r_l\cdot\mathbf{1}+\mathbf{o}(\epsilon)\right) & ,i=1 \\ \tilde{F}^{-1}\mathbf{1} & ,i>1 \end{cases}$

This leads to $\mathcal{A}\left(h_y^S\right) = \mathcal{A}\left(h_y^D\right)$, as required. □

E.14. Proof of claim 12

The shallow fully-connected network considered in this claim is illustrated in fig. 4. Assume ReLU activation $(\sigma(z) = \max\{0,z\})$, and denote by $h_y^{S(fc)}$ the network's y'th score function. We would like to show that $\mathcal{A}(h_y^{S(fc)})$ – the grid tensor of $h_y^{S(fc)}$ w.r.t.the covering templates $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)}$, may take on any value when hidden and output weights ($\{A^z\}_{z\in[Z]}$ and $\mathbf{a}^y$ respectively) are chosen appropriately.

For any $d_1 \ldots d_N \in [M]$, define the following matrix:

$$F^{(d_1\ldots d_N)} := \begin{bmatrix} f_{\theta_1}(\mathbf{x}^{(d_1)}) & \cdots & f_{\theta_M}(\mathbf{x}^{(d_1)}) \\ \vdots & \ddots & \vdots \\ f_{\theta_1}(\mathbf{x}^{(d_N)}) & \cdots & f_{\theta_M}(\mathbf{x}^{(d_N)}) \end{bmatrix} \in \mathbb{R}^{N\times M}$$

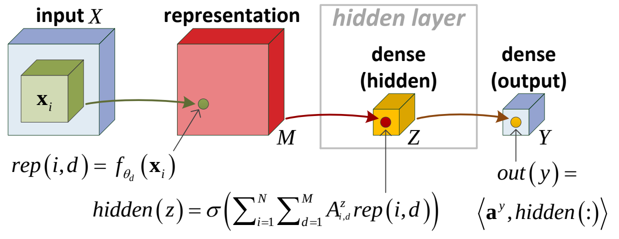
input X
representation
hidden layer
dense (hidden)
dense (output)
$\mathbf{x}_i$
M
Z
Y
$rep(i,d) = f_{\theta_d}(\mathbf{x}_i)$
$hidden(z) = \sigma\left(\sum_{i=1}^{N}\sum_{d=1}^{M} A^z_{i,d}\, rep(i,d)\right)$
$out(y) = \langle \mathbf{a}^y, hidden(:)\rangle$

Figure 4. Shallow fully-connected network obtained by expanding the conv receptive field in the shallow ConvNet to cover the entire spatial extent. The hidden layer consists of a Z-channel dense linear operator weighted by $\{A^z \in \mathbb{R}^{N\times M}\}_{z\in[Z]}$, and followed by point-wise activation $\sigma(\cdot)$. The resulting Z-dimensional vector is mapped to Y network outputs through a dense linear operator weighted by $\{\mathbf{a}^y \in \mathbb{R}^Z\}_{y\in[Y]}$. Best viewed in color.

In words, $F^{(d_1\ldots d_N)}$ is the matrix obtained by taking rows $d_1\ldots d_N$ from F (recurrence allowed), and stacking them one on top of the other. It holds that:

$$\mathcal{A}(h_y^{S(fc)})_{d_1\ldots d_N} = \sum_{z=1}^{Z} a_z^y \max\left\{0, \left\langle F^{(d_1\ldots d_N)}, A^z\right\rangle\right\}$$

where $\langle\cdot,\cdot\rangle$ stands for the inner-product operator, *i.e.* $\left\langle F^{(d_1\ldots d_N)}, A^z\right\rangle := \sum_{i=1}^{N}\sum_{d=1}^{M} F_{i,d}^{(d_1\ldots d_N)} A^z_{i,d}$.

By assumption F has a constant non-zero column. This implies that there exist $j \in [M], c \neq 0$ such that for any $d_1\ldots d_N \in [M]$, all entries in column j of $F^{(d_1\ldots d_N)}$ are equal to c. For every $d_1\ldots d_N \in [M]$ and $z \in [Z]$, denote by $\tilde{F}^{(d_1\ldots d_N)}$ and $\tilde{A}^z$ the matrices obtained by removing the j'th column from $F^{(d_1\ldots d_N)}$ and A^z respectively. Defining $\mathbf{b} \in \mathbb{R}^Z$ to be the vector whose z'th entry is given by $b_z = c\cdot\sum_{i=1}^{N} A^z_{i,j}$, we may write:

$$\mathcal{A}(h_y^{S(fc)})_{d_1\ldots d_N} = \sum_{z=1}^{Z} a_z^y \max\left\{0, \left\langle \tilde{F}^{(d_1\ldots d_N)}, \tilde{A}^z\right\rangle + b_z\right\}$$

noting that for every $z \in [Z]$, $\tilde{A}^z$ and b_z may take on any values with proper choice of A^z. Since by assumption F has non-recurring rows, and since all rows hold the same value (c) in their j'th entry, we have that $\tilde{F}^{(d_1\ldots d_N)} \neq \tilde{F}^{(d'_1\ldots d'_N)}$ for $(d_1\ldots d_N) \neq (d'_1\ldots d'_N)$. An application of lemma 1 now shows that when $Z \geq M^N$, any value for the grid tensor $\mathcal{A}(h_y^{S(fc)})$ may be realized with proper assignment of $\{\tilde{A}^z\}_{z\in[Z]}$, $\mathbf{b}$ and $\mathbf{a}^y$. Since $\{\tilde{A}^z\}_{z\in[Z]}$ and $\mathbf{b}$ may be set arbitrarily through $\{A^z\}_{z\in[Z]}$, we get that with proper choice of hidden and output weights ($\{A^z\}_{z\in[Z]}$ and $\mathbf{a}^y$ respectively), the grid tensor of our network w.r.t.the covering templates may take on any value, precisely meaning that universality holds. □

Lemma 1. *Let $\mathbf{v}_1\ldots\mathbf{v}_k \in \mathbb{R}^D$ be distinct vectors ($\mathbf{v}_i \neq \mathbf{v}_j$ for $i\neq j$), and $c_1\ldots c_k \in \mathbb{R}$ be any scalars. Then, there exist $\mathbf{w}_1\ldots\mathbf{w}_k \in \mathbb{R}^D$, $\mathbf{b} \in \mathbb{R}^k$ and $\mathbf{a} \in \mathbb{R}^k$ such that $\forall i \in [k]$:*

$$\sum_{j=1}^{k} a_j \max\{0, \mathbf{w}_j^\top \mathbf{v}_i + b_j\} = c_i \tag{15}$$

Proof. As shown in the proof of claim 1, for distinct $\mathbf{v}_1\ldots\mathbf{v}_k \in \mathbb{R}^D$ there exists a vector $\mathbf{u} \in \mathbb{R}^D$ such that $\mathbf{u}^\top\mathbf{v}_i \neq \mathbf{u}^\top\mathbf{v}_j$ for all $1\leq i < j \leq k$. We assume without loss of generality that $\mathbf{u}^\top\mathbf{v}_1 < \ldots < \mathbf{u}^\top\mathbf{v}_k$, and set $\mathbf{w}_1\ldots\mathbf{w}_k$, $\mathbf{b}$ and $\mathbf{a}$ as follows:

- $\mathbf{w}_1 = \cdots = \mathbf{w}_k = \mathbf{u}$
- $b_1 = -\mathbf{u}^\top\mathbf{v}_1 + 1$
- $b_j = -\mathbf{u}^\top\mathbf{v}_{j-1}$ for $j = 2\ldots k$
- $a_1 = c_1$
- $a_j = \frac{c_j - c_{j-1}}{\mathbf{u}^\top\mathbf{v}_j - \mathbf{u}^\top\mathbf{v}_{j-1}} - \sum_{t=1}^{j-1} a_t$ for $j = 2\ldots k$

To complete the proof, we show below that this assignment meets the condition in eq. 15 for $i = 1\ldots k$.

The fact that:

$$\mathbf{w}_j^\top\mathbf{v}_1 + b_j = \begin{cases} \mathbf{u}^\top\mathbf{v}_1 - \mathbf{u}^\top\mathbf{v}_1 + 1 = 1 & , j = 1 \\ \mathbf{u}^\top\mathbf{v}_1 - \mathbf{u}^\top\mathbf{v}_{j-1} \leq 0 & , 2\leq j\leq k \end{cases}$$

implies that the condition in eq. 15 indeed holds for $i = 1$:

$$\sum_{j=1}^{k} a_j \max\{0, \mathbf{w}_j^\top\mathbf{v}_1 + b_j\} = a_1\cdot 1 + \sum_{j=1}^{k} a_j\cdot 0 = a_1 = c_1$$

For $i > 1$ we have:

$$\mathbf{w}_j^\top\mathbf{v}_i + b_j = \begin{cases} \mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_1 + 1 > 0 & , j = 1 \\ \mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_{j-1} > 0 & , 2\leq j\leq i \\ \mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_{j-1} \leq 0 & , i < j\leq k \end{cases}$$

which implies:

$$\sum_{j=1}^{k} a_j \max\{0, \mathbf{w}_j^\top\mathbf{v}_i + b_j\} = \\ a_1(\mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_1 + 1) + \sum_{j=2}^{i} a_j(\mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_{j-1})$$

Comparing this to the same expression with i replaced by $i-1$ we obtain:

$$\sum_{j=1}^{k} a_j \max\{0, \mathbf{w}_j^\top\mathbf{v}_i + b_j\} = \\ \sum_{j=1}^{k} a_j \max\{0, \mathbf{w}_j^\top\mathbf{v}_{i-1} + b_j\} + \\ (\mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_{i-1})\sum_{j=1}^{i} a_j$$

Now, if we follow an inductive argument and assume that the condition in eq. 15 holds for $i-1$, *i.e.*that $\sum_{j=1}^{k} a_j \max\{0, \mathbf{w}_j^\top\mathbf{v}_{i-1} + b_j\} = c_{i-1}$, we get:

$$\sum_{j=1}^{k} a_j \max\{0, \mathbf{w}_j^\top\mathbf{v}_i + b_j\} = \\ c_{i-1} + (\mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_{i-1})\sum_{j=1}^{i} a_j$$

Plugging in the definition $a_i = \frac{c_i - c_{i-1}}{\mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_{i-1}} - \sum_{j=1}^{i-1} a_j$ gives:

$$\sum_{j=1}^{k} a_j \max\{0, \mathbf{w}_j^\top\mathbf{v}_i + b_j\} = \\ c_{i-1} + (\mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_{i-1})\frac{c_i - c_{i-1}}{\mathbf{u}^\top\mathbf{v}_i - \mathbf{u}^\top\mathbf{v}_{i-1}} = c_i$$

Thus the condition in eq. 15 holds for i as well. We have therefore shown by induction that our assignment of $\mathbf{w}_1\ldots\mathbf{w}_k$, $\mathbf{b}$ and $\mathbf{a}$ meets the lemma's requirement. □

E.15. Proof of claim 13

The proof here is almost identical to that of claim 7. The only difference is that in the latter, we used the fact that the generalized HT decomposition (eq. 7), when equipped with $g(a,b) = a{\cdot}b$, almost always produces tensors whose matrix arrangements have rank at least $\min\{r_0, M\}^{N/2}$, whereas here, we require an analogous result for the *shared* generalized HT decomposition (eq. 9). Such result is provided by the proof of theorem 1 in (Cohen et al., 2016b). □

E.16. Proof of claim 14

In the proof of claim 8 we have shown, for arbitrary distinct templates $\mathbf{x}^{(1)} \ldots \mathbf{x}^{(M)} \in \mathbb{R}^s$, an explicit weight setting for the deep ConvNet with ReLU activation and max pooling, leading the latter to produce a grid tensor that cannot be realized by a shallow ConvNet with ReLU activation and max pooling, if that has less than $\min\{r_0, M\}^{N/2} \cdot \frac{2}{M \cdot N}$ hidden channels. Since the given weight setting was location invariant, *i.e.*the assignment of $\mathbf{a}^{l,j,\gamma}$ did not depend on j, it applies as is to a deep ConvNet with weight sharing, and the desired result readily follows. □

E.17. Proof of claim 15

The proof is similar in spirit to that of claim 9, which dealt with incompleteness of depth efficiency under ReLU activation and max pooling in the unshared setting. Our focus here is on the shared setting, or more specifically, on the case where the deep ConvNet is limited by weight sharing while the shallow ConvNet is not. Accordingly, we would like to show the following. Fixing $g(a,b) = \max\{a,b,0\}$, per arbitrary invertible F there exists a weight ($\mathbf{a}^{l,\gamma}$) setting for the shared generalized HT decomposition (eq. 9), such that the produced tensor may be realized by the generalized CP decomposition (eq. 6) with $Z = 1$, and this holds even if the weights $\mathbf{a}^{l,\gamma}$ and matrix F are subject to small perturbations.

Before heading on to prove that a weight setting as above exists, we introduce a new definition that will greatly simplify our proof. We refer to a tensor $\mathcal{A}$ of order P and dimension M in each mode as *basic*, if there exists a vector $\mathbf{u} \in \mathbb{R}^M$ with non-decreasing entries ($u_1 \leq \ldots \leq u_M$), such that $\mathcal{A} = \mathbf{u} \otimes_g \cdots \otimes_g \mathbf{u}$ (*i.e.*$\mathcal{A}$ is equal to the generalized tensor product of $\mathbf{u}$ with itself P times, with underlying operation $g(a,b) = \max\{a,b,0\}$). A basic tensor can obviously be realized by the generalized CP decomposition (eq. 6) with $Z = 1$ (given that non-degeneracy is used to ensure the latter's representation matrix is non-singular), and so it suffices to find a weight ($\mathbf{a}^{l,\gamma}$) setting for the shared generalized HT decomposition (eq. 9) that gives rise to a basic tensor, and in addition, ensures that small perturbations on the weights $\mathbf{a}^{l,\gamma}$ and matrix F still yield basic tensors. Two trivial facts that relate to basic tensors and will be used in our proof are: (i) the generalized tensor product of a basic tensor with itself is basic, and (ii) a linear combination of basic tensors with non-negative weights is basic.

Turning to the main part of the proof, we now show that the following weight setting meets our requirement:

- $\mathbf{a}^{0,\gamma} = F^{-1}\mathbf{v}$
- $\mathbf{a}^{l,\gamma} = \mathbf{1},\ l \in [L-1]$
- $\mathbf{a}^{L,y} = \mathbf{1}$

$\mathbf{v}$ here stands for the vector $[1,2,\ldots,M]^\top \in \mathbb{R}^M$, and $\mathbf{1}$ is an all-1 vector with dimension to be understood by context. Let $\mathcal{E}^F$ be an additive noise matrix applied to F, and $\{\boldsymbol{\epsilon}^{l,\gamma}\}_{l,\gamma}$ be additive noise vectors applied to $\{\mathbf{a}^{l,\gamma}\}_{l,\gamma}$. We would like to prove that under the weight setting above, when applied noise ($\mathcal{E}^F, \boldsymbol{\epsilon}^{l,\gamma}$) is small enough, the grid tensor produced by the shared generalized HT decomposition (eq. 9) is basic.

For convenience, we adopt the notation $\mathbf{o}(\epsilon)$ as referring to vectors that tend to $\mathbf{0}$ as $\mathcal{E}^F \to 0$ and $\boldsymbol{\epsilon}^{l,\gamma} \to \mathbf{0}$, with the dimension of a vector to be understood by context. Plugging in the noisy variables into the shared generalized HT decomposition (eq. 9), we get for every $\alpha \in [r_0]$:

$$\begin{aligned} &((F+\mathcal{E}^F)(\mathbf{a}^{0,\alpha}+\boldsymbol{\epsilon}^{0,\alpha})) \otimes_g ((F+\mathcal{E}^F)(\mathbf{a}^{0,\alpha}+\boldsymbol{\epsilon}^{0,\alpha})) \\ &= ((F+\mathcal{E}^F)(F^{-1}\mathbf{v}+\boldsymbol{\epsilon}^{0,\alpha})) \otimes_g ((F+\mathcal{E}^F)(F^{-1}\mathbf{v}+\boldsymbol{\epsilon}^{0,\alpha})) \\ &= \tilde{\mathbf{v}}^\alpha \otimes_g \tilde{\mathbf{v}}^\alpha \end{aligned}$$

where $\tilde{\mathbf{v}}^\alpha = \mathbf{v} + \mathbf{o}(\epsilon)$. If the applied noise ($\mathcal{E}^F, \boldsymbol{\epsilon}^{l,\gamma}$) is small enough the entries of $\tilde{\mathbf{v}}^\alpha$ are non-decreasing and $\tilde{\mathbf{v}}^\alpha \otimes_g \tilde{\mathbf{v}}^\alpha$ is a basic tensor (matrix). Moving to the next level of the decomposition, we have for every $\gamma \in [r_1]$:

$$\phi^{1,\gamma} = \sum_{\alpha=1}^{r_0} (a_\alpha^{1,\gamma} + \epsilon_\alpha^{1,\gamma}) \cdot \tilde{\mathbf{v}}^\alpha \otimes_g \tilde{\mathbf{v}}^\alpha$$

When applied noise ($\mathcal{E}^F, \boldsymbol{\epsilon}^{l,\gamma}$) is small enough the weights of this linear combination are non-negative, and together with the tensors (matrices) $\tilde{\mathbf{v}}^\alpha \otimes_g \tilde{\mathbf{v}}^\alpha$ being basic, this leads $\phi^{1,\gamma}$ to be basic as well. Continuing in this fashion over the levels of the decomposition, we get that with small enough noise, for every $l \in [L-1]$ and $\gamma \in [r_l]$, $\phi^{l,\gamma}$ is a basic tensor. A final step in this direction shows that under small noise, the produced grid tensor $\mathcal{A}\left(h_y^D\right)$ is basic as well. This is what we set out to prove. □

F. On the Incidence of Depth Efficiency

In claim 7 we saw that depth efficiency is complete with linear activation and product pooling. That is to say, with linear activation and product pooling, besides a negligible set, all weight settings for the deep ConvNet (fig. 1 with size-2 pooling windows and $L = \log_2 N$ hidden layers) lead to score functions that cannot be realized by the shallow ConvNet (fig. 2) unless the latter has super-polynomial size. We have also seen (claims 8 and 9) that replacing the activation and pooling operators by ReLU and max respectively, makes depth efficiency incomplete. There are still weight settings leading the deep ConvNet to generate score functions that require the shallow ConvNet to have super-polynomial size, but these do not occupy the entire space. In other words, there is now positive measure to the set of deep ConvNet weight configurations leading to score functions efficiently realizable by the shallow ConvNet. A natural question would then be just how frequent depth efficiency is under ReLU activation and max pooling. More formally, we may consider a uniform distribution over a compact domain in the deep ConvNet's weight space, and ask the following. Assuming weights for the deep ConvNet are drawn from this distribution, what is the probability that generated score functions exhibit depth efficiency, *i.e.*require super-polynomial size from the shallow ConvNet? In this appendix we address this question, arguing that the probability tends to 1 as the number of channels in the hidden layers of the deep ConvNet grows. We do not prove this formally, but nonetheless provide a framework

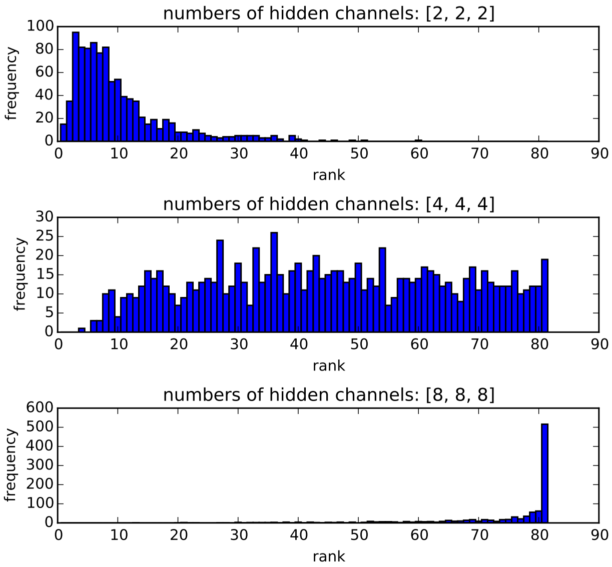

Figure 5. Simulation results demonstrating that under ReLU activation and max pooling, the incidence of depth efficiency increases as the number of channels in the hidden layers of the deep ConvNet ($r_0 \ldots r_{L-1}$) grows. The plots show histograms of the ranks produced by the matricized generalized HT decomposition (eq. 10) with $g(a,b) = \max\{a,b,0\}$. The number of levels in the decomposition was set to $L = 3$ (implying input size of $N = 2^L = 8$). The size of the representation matrix F was set through $M = 3$, and the matrix itself was fixed to the identity. Weights ($\mathbf{a}^{l,j,\gamma}$) were drawn at random independently and uniformly from the interval $[-1,1]$. Three channel-width configurations were tried: (i) $r_0 = r_1 = r_2 = 2$ (ii) $r_0 = r_1 = r_2 = 4$ (ii) $r_0 = r_1 = r_2 = 8$. For each configuration 1000 random tests were run, creating the histograms presented in the figure (each test produced a single matrix $[\mathcal{A}(h_y^D)]$, accounting for a single entry in a histogram). As can be seen, the distribution of the produced rank ($rank[\mathcal{A}(h_y^D)]$) tends towards the maximum ($M^{N/2} = 81$) as the numbers of hidden channels grow.

we believe may serve as a basis for establishing formal results concerning the incidence of depth efficiency. The framework, which relies on the material delivered in app. E.2, is not limited to ReLU activation and max pooling – it may be used under different choices of activation and pooling operators as well.

The central tool used in the paper for proving depth efficiency is the rank of grid tensors when these are arranged as matrices. We establish upper bounds on the rank of matricized grid tensors produced by the shallow ConvNet through the matricized generalized CP decomposition (eq. 11). These upper bounds are typically linear in the size of the input (N) and the number of hidden channels in the network (Z). The challenge is then to derive a super-polynomial (in N) lower bound on the rank of matricized grid tensors produced by the deep ConvNet through the matricized generalized HT decomposition (eq. 10). In the case of linear activation and product pooling ($g(a,b) = a{\cdot}b$), the generalized Kronecker product $\odot_g$ reduces to the standard Kronecker product $\odot$, and the rank-multiplicative property of the latter ($rank(A \odot B) = rank(A){\cdot}rank(B)$) can be used to show (see (Cohen et al., 2016b)) that besides in negligible (zero measure) cases, rank grows rapidly through the levels of the matricized generalized HT decomposition (eq. 10), to the point where the final produced matrix has exponential rank. This situation does not persist when the activation and pooling operators are replaced by ReLU and max (respectively). Indeed, in the proof of claim 9 we explicitly presented a non-negligible (positive measure) case where the matricized generalized HT decomposition (eq. 10) produces a matrix of rank 1. To study the incidence of depth efficiency under ReLU activation and max pooling, we assume the weights ($\mathbf{a}^{l,j,\gamma}$) of the matricized generalized HT decomposition (eq. 10) are drawn independently and uniformly from a bounded interval (*e.g.* $[-1,1]$), and question the probability of the produced matrix $[\mathcal{A}\left(h_y^D\right)]$ having rank super-polynomial in N.

To study $rank[\mathcal{A}\left(h_y^D\right)]$, we sequentially traverse through the levels $l = 1 \ldots L$ of the matricized generalized HT decomposition (eq. 10), at each level going over all locations $j \in [N/2^l]$. When at location j of level l, for each $\alpha \in [r_{l-1}]$, we draw the weights $\mathbf{a}^{l-1,2j-1,\alpha}$ and $\mathbf{a}^{l-1,2j,\alpha}$ (independently of the previously drawn weights), and observe the random variable $R^{l,j,\alpha}$, defined as the rank of the matrix $[\phi^{l-1,2j-1,\alpha}] \odot_g [\phi^{l-1,2j,\alpha}]$. Given the weights drawn while traversing through the previous levels of the decomposition, the random variables $\{R^{l,j,\alpha} \in \mathbb{N}\}_{\alpha \in [r_{l-1}]}$ are independent and identically distributed. The random variable $R^{l,j} := \max_{\alpha \in [r_{l-1}]}\{R^{l,j,\alpha}\}$ thus tends to concentrate on higher and higher values as r_{l-1} (number of channels in hidden layer $l-1$ of the deep ConvNet) grows. When the next level ($l+1$) of the decomposition will be traversed, the weights $\{\mathbf{a}^{l,j,\gamma}\}_{\gamma \in [r_l]}$ will be drawn, and the matrices $\{[\phi^{l,j,\gamma}]\}_{\gamma \in [r_l]}$ will be generated. According to claim 16 below, with probability 1, all of these matrices will have rank equal to at least $R^{l,j}$. We conclude that, assuming the generalized Kronecker product $\odot_g$ has the potential of increasing the rank of its operands, ranks will generally ascend across the levels of the matricized generalized HT decomposition (eq. 10), with steeper ascends being more and more probable as the number of channels in the hidden layers of the deep ConvNet ($r_0 \ldots r_{L-1}$) grows.

The main piece that is missing in order to complete the sketch we have outlined above into a formal proof, is the behavior of rank under the generalized Kronecker product $\odot_g$. This obviously depends on the choice of underlying operator g. In the case of linear activation and product pooling $g(a,b) = a{\cdot}b$, the generalized Kronecker product $\odot_g$ reduces to the standard Kronecker product $\odot$, and ranks always increase multiplicatively, *i.e.* $rank(A \odot B) = rank(A){\cdot}rank(B)$ for any matrices A and B. The fact that there is a simple law governing the behavior of ranks makes this case relatively simple to analyze, and we indeed have a full characterization (claim 7). In the case of *linear* activation and max pooling the underlying operator is given by $g(a,b) = \max\{a,b\}$, and it is not difficult to see that $\odot_g$ does not decrease rank, *i.e.*$rank(A \odot_g B) \geq \min\{rank(A), rank(B)\}$ for any matrices A and B [9]. For ReLU activation and max pooling, corresponding to the choice $g(a,b) = \max\{a,b,0\}$, there is no simple rule depicting the behavior of ranks under $\odot_g$, and in fact, for matrices A and B holding negative values, the rank of $rank(A \odot_g B)$ necessarily drops to zero. Nonetheless, it seems reasonable to assume that at least in some cases, a non-linear operation such as $\odot_g$ does

[9] To see this, simply note that under the choice $g(a,b) = \max\{a,b\}$ there is either a sub-matrix of $A \odot_g B$ that is equal to A, or one that is equal to B.

increase rank, and as we have seen, benefiting from these cases is more probable when the hidden layers of the deep ConvNet include many channels. To this end, we provide in fig. 5 simulation results for the case of ReLU activation and max pooling ($g(a,b) = \max\{a,b,0\}$), demonstrating that indeed ranks produced by the matricized generalized HT decomposition (eq. 10) tend to be higher as $r_0 \ldots r_{L-1}$ grow. We leave a complete formal analysis of this phenomenon to future work.

Claim 16. *Let $A_1 \ldots A_m$ be given matrices of the same size, having ranks $r_1 \ldots r_m$ respectively. For every weight vector $\boldsymbol{\alpha} \in \mathbb{R}^m$ define the matrix $A(\boldsymbol{\alpha}) := \sum_{i=1}^{m} \alpha_i A_i$, and suppose we randomize $\boldsymbol{\alpha}$ by some continuous distribution. Then, with probability 1, we obtain a matrix $A(\boldsymbol{\alpha})$ having rank at least $\max_{i\in[m]} r_i$.*

Proof. Our proof relies on concepts and results from Lebesgue measure theory (see app. E.1 for a brief discussion). The result to prove is equivalent to stating that there is measure zero to the set of weight vectors $\boldsymbol{\alpha}$ for which $rank(A(\boldsymbol{\alpha})) < \max_{i\in[m]} r_i$.

Assume without loss of generality that $\max_{i\in[m]} r_i$ is equal to r_1, and that the top-left $r_1 \times r_1$ block of A_1 is non-singular. For every $\boldsymbol{\alpha}$ define $p(\boldsymbol{\alpha}) := \det(A(\boldsymbol{\alpha})_{1:r_1,1:r_1})$, *i.e.*$p(\boldsymbol{\alpha})$ is the determinant of the $r_1 \times r_1$ top-left block of the matrix $A(\boldsymbol{\alpha})$. $p(\boldsymbol{\alpha})$ is obviously a polynomial in the entries of $\boldsymbol{\alpha}$, and by assumption $p(\mathbf{e}_1) \neq 0$, where $\mathbf{e}_1 \in \mathbb{R}^m$ is the vector holding 1 in its first entry and 0 elsewhere. Since a non-zero polynomial vanishes only on a set of zero measure (see (Caron and Traynor, 2005) for example), the set of weight vectors $\boldsymbol{\alpha}$ for which $p(\boldsymbol{\alpha}) = 0$ has measure zero. This implies that the top-left $r_1 \times r_1$ block of $A(\boldsymbol{\alpha})$ is non-singular almost everywhere, and in particular $rank(A(\boldsymbol{\alpha})) \geq r_1 = \max_{i\in[m]} r_i$ almost everywhere. □